異なる統計モデルを用いたニュース記事からの重要箇所の抽出

皿海　宏明†　　湯本　高行††

† 兵庫県立大学大学院工学研究科　〒 671-2201 兵庫県姫路市書写 2167
†† 兵庫県立大学大学院工学研究科　〒 671-2201 兵庫県姫路市書写 2167
E-mail: †ei16c004@steng.u-hyogo.ac.jp, ††yumoto@eng.u-hyogo.ac.jp

あらまし　ニュース記事から重要箇所を抽出するために異なる統計モデルを用いて語の単位での重要かどうかを推定する手法を提案する．本論文では重要かどうかを抜粋のされやすさとし，SVM を用いて異なる 3 つの統計モデルを特徴量とし重要かどうかの推定を行う．モデルには DF を用いたコレクションにおける出現確率が低いものを高く評価するモデル，LDA を用いた入力文章のトピックに沿った語を高く評価するモデル，格フレームを用いた述語に対して頻出する項の格の語を高く評価するモデルを使用する．TF-IDF を用いるモデルをベースラインとした評価実験の結果，本手法の有用性を示すことができた．

キーワード　テキストマイニング, 文書要約, LDA, 格フレーム, SVM

1. はじめに

近年，インターネットによるニュースの提供は広く普及している．ユーザーは情報の取捨選択を行うことが求められている．ユーザーがネットのニュース記事の情報を効率よく扱う技術の一つとして重要箇所の抽出がある．本研究は重要箇所を抽出することにより、ユーザーのテキストの把握を容易にする．本論文では抜粋されやすい語は重要であると仮定し，重要かどうかの定義を抜粋のされやすさとする．重要箇所かどうかは人手によって抜粋されたかどうかで決定する．文節単位での重要かどうかの推定は，文単位での重要かどうかの推定より含まれる形態素の手がかりが少ないため，単純な頻度を重要視する手法では精度が悪くなると考えられる．そこで文節の持つ格フレームを手がかりとするアプローチを入れることによってよりよい推定を行えるのではないかと考える．また，記事にはトピックが存在するとしたとき，そのトピックに頻出する語というのは重要かどうかの推定の手がかりになるのではないかと考える．文単位での TFIDF モデルを用いて記事中の頻度とコーパス上の逆頻度より重要かどうかを推定する従来手法と比べ，本手法では文節単位でのコーパス上の語の出現確率モデル (DF モデル)，記事のトピックごとの語の出現確率 (LDA モデル)，述語に対する項の格の共起確率 (格フレームモデル) より重要かどうかを決定することで，より多角的な統計的確率を元に重要かどうかの推定を行う．提案した統計モデルの指標をより推定に反映させるため，重要かどうかの推定は各モデルの指標を特徴量として SVM [1] を用いて行う．

2. 関連研究

重要箇所を抜粋されやすい語としているため，文書要約を関連としている．Zechner [2] の研究では文の重要かどうかを含まれる語の TFIDF の重みの総和とする重要文抽出を行っている．本手法のベースラインはこの研究を参考にした．堀内ら [3] の研究では IDF と N-gram を用いた文短縮を行っている．意味的な重要箇所の評価が可能な IDF と日本語らしさの評価が可能な統計的言語モデルの N-gram を組み合わせることで文を不自然でない意味を持った文に短縮している．本研究では重要箇所の抽出に DF を用いたモデルを使用している．畑山ら [4] の研究では格フレーム辞書を用いて語句単位での重要箇所を抽出し，再構成して要約文を生成している．本研究でも重要箇所の抽出に格フレームを使用しているが，本研究では格の種類ごとの出現確率に注目している点で主語や目的語の抽出を行っているこの研究と異なる．

3. 提案手法

本研究では，重要箇所を抽出する問題を，文節を重要か否かの 2 値に分類する問題として扱う．我々は文節の統計的な重要性について以下の 3 つの観点に注目する．まず，1 つ目はコーパス全体での語の出現状況である．2 つ目は関連するトピックにおける語の出現状況である．3 つ目は文法上の共起関係である．本研究では述語と項の共起関係に注目する．これらは表 1 に示すようにそれぞれ重要視する点が異なるため，どれか 1 つを選択するよりも組み合わせて使用することが有効であると考えている．そこで本研究では，それぞれの観点に基づく指標を素性とする SVM により分類器を構築する．

表 1　各モデルの特徴と重要視する点

モデル	特徴	重要視する点
DF	入力文章に依存せず，一般語を判断する．	希少性
LDA	入力文章のトピック語を判断する．	トピックへの関連性
格フレーム	述語と項の関係において頻出する語を判断する．	述語にとっての必要性

3.1 DF を用いたモデル

DF を用いたモデルによってコレクションに頻出かどうかを求める．DF を用いたモデルによる手法を示す．コーパスを形

態素解析し，記事ごとの動詞，名詞，形容詞，副詞を文書コレクション C とする．入力文を形態素解析し，動詞，名詞，形容詞，副詞を入力語集合 S とし，含まれる語 w についてそれぞれ指標を算出する．複合語など，文節に対象となる複数の語が含まれる場合は，含まれる語の評価値を，抜粋時に用いた文節単位で積をとる．語 w についての DF を用いたモデルの指標 $Score_{DF}(w)$ を求める式を以下に示す．

$$Score_{DF}(w) = P(w|C) = \frac{DF(w,C)}{|C|}$$

$P(w|C)$ は語 w が文書コレクション C の文書に出現する頻度を C に含まれる文書数で割ったものである．$DF(w,C)$ は語 w が文書コレクション C の文書に出現する頻度である．$|C|$ は C の文書数である．

3.2 LDA を用いたモデル

LDA [5] を用いたモデルによって入力文章のトピックに沿った語であるかどうかを求める．LDA を用いたモデルによる手法を示す．コーパスを形態素解析し，記事ごとに動詞，名詞，形容詞，副詞を文書コレクション C とする．これより LDA モデルを作成し，形態素の出現確率をトピック集合 T の要素であるトピック t に対してそれぞれ求め，$P(w|t)$ とする．入力文を形態素解析し，動詞，名詞，形容詞，副詞を入力語集合 S とし，含まれる語 w についてそれぞれ指標を算出する．文節に対象となる複数の語が含まれる場合は，含まれる語の指標を，抜粋時に用いた文節単位で積をとる．語 w についての LDA を用いたモデルの指標 $Score_{LDA}(w)$ を求める式を以下に示す．

$$Score_{LDA}(w) = P(w|t)$$

$P(w|t)$ は事前に求めているトピック t における語 w の出現確率とする．入力語集合をクエリとみなしたときのトピックに対するクエリ尤度 $P(t|S)$ に基づいて最も適切なトピックと思われる t を選択する．すなわち， $P(t|S)$ が最大となる t を選ぶ．

$$P(t|S) = \prod_{w \in W} \left((1-\lambda-\alpha)P(w|t) + \lambda \frac{\sum_{t' \in T} P(w|t')}{|T|} + \alpha \right)$$

$P(t|S)$ のクエリ尤度を求める場合はパラメータ λ， α によってマクロ平均スムージング，定数値によるスムージングをおこなう．マクロ平均スムージングだけでは，入力語集合に文書コレクション中に一度も出現しない語が一つでも含まれていると，クエリ尤度の値がすべてのトピックに対して 0 になってしまうため，定数値によるスムージングを導入している．

LDA を用いたモデルの例を示す．入力文章が「横綱」「場所」「相撲」「震災」からなる A と「横綱」「原発」「稼働」「震災」からなる B があるとする．語の比率から A は相撲トピックの文章であり，B は原発震災トピックの文章であることが分かる．この二つの入力文章に対して LDA の指標をとったものを表 2 に示す．

表 2 より，相撲トピックである A では異なるトピックの「震災」が指標 0 に，原発震災トピックである B では異なるトピックの「横綱」が指標 0 となっている．また,「震災」は自身のトピックである B では指標が高くなり,「横綱」も自身のトピックである A では指標が高くなっている．以上より，LDA を用いたモデルの指標は入力文章のトピックに依存し，入力文章と同じトピックの語は高い指標がつくことが分かる．

表 2　入力文章のトピックごとの LDA モデルの指標の違い

	A				B			
語	横綱	場所	相撲	震災	横綱	原発	稼働	震災
指標	0.012	0.031	0.009	0.000	0.000	0.033	0.011	0.011

3.3 格フレームを用いたモデル

格フレームを用いたモデルによって述語に対して頻繁に共起する格であるかを求める．格フレームを用いたモデルによる手法を示す．コーパスを格解析し，述語について述語項関係で出現する項を格ごとにまとめ，格フレーム辞書とする．入力文を形態素解析し，動詞，名詞，形容詞，副詞を入力語集合 S とし，含まれる語 w についてそれぞれ指標を算出する．文節に対象となる複数の語が含まれる場合は，含まれる語の指標を抜粋時に用いた文節単位で積をとる．w が述語のときに対応する項集合を $SEC(w)$ とする．w が項のときに対応する述語を $pred(w)$ とし，w の格は $cf(w)$ と表す．語 w についての格フレームを用いたモデルの指標 $Score_{CFC}(w)$ を求める式を以下に示す．なお，述語に対して項の格が一意に決定するような述語，項は述語が一般的なものでないことが多いため考慮しない．

$$Score_{CFC}(w) = \begin{cases} P(cf(w)|pred(w)) & (w \text{ が項}) \\ P(cf(SEC(w))|w) & (w \text{ が述語}) \end{cases}$$

$$P(cf(SEC(w))|w) = \prod_{w' \in SEC(w)} P(cf(w')|w)$$

$P(cf(w)|pred(w))$ は項 w と対応する述語 $pred(w)$ における w の格の出現確率である．$P(cf(SEC(w))|w)$ は w における述語 w と対応する項集合 $SEC(w)$ の格の尤度である．

格フレームを用いたモデルの例を示す．入力文章を「郵送で食品が届く」とする．格解析を行うと,「郵送で」はデ格,「食品が」はガ格,「届く」は述語だと分かる．それぞれの文節で格フレームを用いたモデルによる指標を求める.「届く」におけるデ格の出現確率は 0.04，ガ格の出現確率は 0.37 とする．

$$\begin{aligned} Score_{CFC}(\text{郵送}) &= P(\text{デ格} \mid \text{届く}) = 0.04 \\ Score_{CFC}(\text{食品}) &= P(\text{ガ格} \mid \text{届く}) = 0.37 \\ Score_{CFC}(\text{届く}) &= P(\{ \text{デ格}, \text{ガ格} \} \mid \text{届く}) \\ &= P(\text{デ格} \mid \text{届く}) \times P(\text{ガ格} \mid \text{届く}) = 0.01 \end{aligned}$$

式より,「届く」に対する出現率の高いガ格の「食品が」がデ格の「郵送で」より指標が高い．このことは「届く」にとって「郵送」よりも「食品」の方が重要であることを意味する。

3.4 SVM を用いたモデルの統合と重要性の推定

DF，LDA，格フレームを用いたモデルによって求めた指標より重要かどうかの推定を行う．これらの指標に最適な重みを与えて重要かどうかを推定するために SVM を使用する．文節

ごとの各モデルの指標を特徴量，人手によって抜粋されたかどうかをクラスラベルとして SVM を用いて推定を行う．

4. 評価実験

4.1 実験方法

入力文章の文節に対して，DF，LDA，格フレームを用いたモデルによってそれぞれ指標を求める．コーパスは「CD-毎日新聞データ集 2014 年版」と 2014 年版の「読売新聞記事データ」のタイトルを除く本文のみ 10 万件とする．形態素解析には JUMAN，格解析には KNP，LDA モデルには Python の gensim ライブラリを使用する．

格フレーム辞書は，KNP によって「CD-毎日新聞データ集 2014 年版」と 2014 年版の「読売新聞記事データ」のタイトルを除く本文のみ 1 万件からなるコーパスを格解析し，述語と述語項関係を持つ項とその格を述語ごとにまとめることで構築する．格フレームモデルでは述語に対する項の格の出現確率を格フレーム辞書より参照する．

入力文章は新聞のリード 20 記事とする．LDA モデルのトピック数 $|T|$ は 50，λ は 0.2，α は 0.001 とする．入力文章の文節に対して，人手で 50 %程度の抜粋を行い，抜粋されたかどうかをクラスラベルとする．モデルの指標を特徴量として，データセットを作成する．LIBSVM により 10 分割交差検定を行う．グリッドサーチによりパラメータを決定した．ベースラインとして TFIDF モデルを用いて同様にする．個々のモデルの評価のためにモデルをそれぞれ単独で特徴量として使用した場合も同様にする．

4.2 TFIDF を用いたモデル

ベースラインの指標となる TFIDF を用いたモデルによる手法を示す．コーパスを形態素解析し，記事ごとに動詞，名詞，形容詞，副詞を文書コレクション C とする．入力文を形態素解析し，動詞，名詞，形容詞，副詞を入力語集合 W とし，含まれる語 w についてそれぞれ指標を算出する文節に対象となる複数の語が含まれる場合は，含まれる語の指標を，抜粋時に用いた文節単位で相加平均をとる．語 w についての TFIDF を用いたモデルの指標 $Score_{TFIDF}(w)$ を求める式を以下に示す．

$$Score_{TFIDF}(w) = P(w|W) \times log\frac{1}{P(w|C)}$$

$P(w|W)$ は入力語集合 W 中で w が出現する確率である．$P(w|C)$ は w が文書コレクション C の文書に出現する頻度を C に含まれる文書数で割ったものである．

4.3 実験結果

表 3 に 10 分割交差検定の正解率の平均値，グリッドサーチにより決定された SVM のパラメータ gamma，cost を示す．図 4 に 10 分割交差検定の正解率の箱ひげ図を示す．また，提案手法と TFIDF について詳細に比較するため図 2 に分割ごとの正解率の散布図を示す．

表 3 より LDA と提案手法の正解率が高く，ベースラインである TFIDF モデルと比較して抜粋された語をよりよく推定していることが分かる．また提案手法を構成する統計モデル 3 つはどれもベースラインを正解率の平均で超えていないので，組み合わせることで互いが補完的に推定を助けていることが分かる．

表 3　提案手法及び各モデルの正解率とパラメータ

	提案手法	DF	LDA	CFC	TFIDF
gamma	10^{13}	10^{15}	10^{10}	10^{14}	10^{12}
cost	10	1	1	1	1
accuracy	0.741	0.701	0.700	0.705	0.726

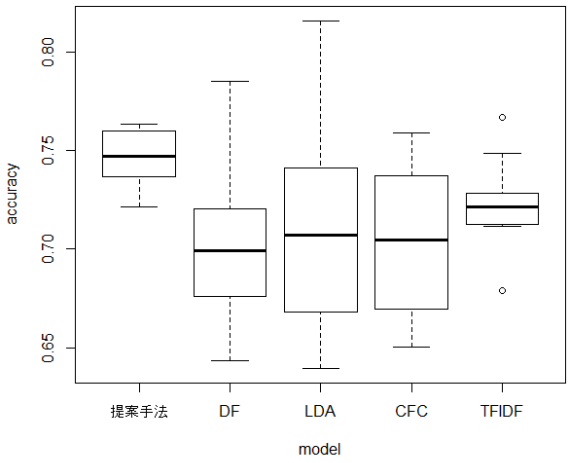

図 1　10 分割交差検定の正解率の箱ひげ図

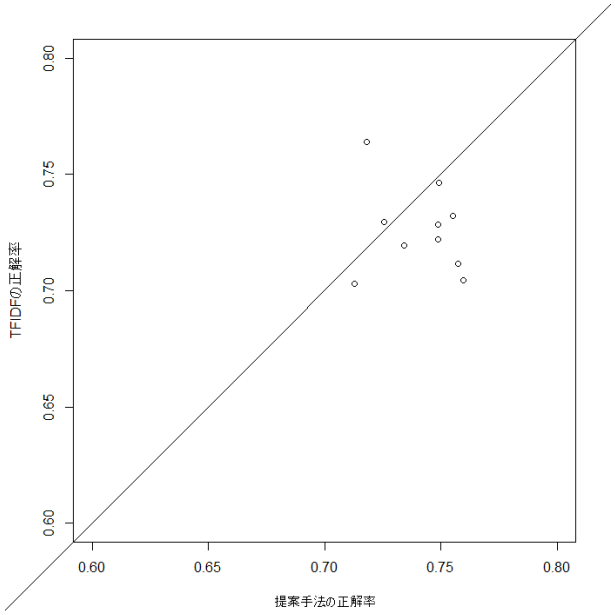

図 2　提案手法と TFIDF の正解率の散布図

図 4 より提案手法はベースラインである TFIDF と比較して正解率の中央値が高いことが分かる．また提案手法と比較して DF，LDA，格フレーム (CFC) を用いたモデルは正解率のばらつきが大きいことがわかる．

図 2 より提案手法と LDA モデルを用いた手法の性能を比較する．提案手法と TFIDF の正解率の散布図の $y = x$ の補助

線より上の点は TFIDF モデルの方が提案手法より重要かどうかの推定ができているデータである．補助線の上より下にあるデータの数が多く，提案手法が優れていると言える．

4.4 考　　察

統計モデルを組み合わせた手法を提案し，ベースラインの性能を超えることは確認できた．しかし，統計モデル単体の精度が低い．これは指標が確率的な値をそのまま用いており確率分布の冪乗則などが考慮されていないためだと考えられる．LDA においては，具体的な指標の高い語を調べると，頻出する意味を持たない不要な語が多かった．これでは単体でトピックに関連する語を適切に区別できない．しかし純粋に頻度のみをみる DF の手法と組み合わせて差分からトピック語を区別することが可能になると考えられる．

5. 具　体　例

新聞記事のリード 1 つに対して評価実験と同様に DF，LDA，格フレームを用いたモデルの指標を求め，傾向を比較する．各モデルごとにどのような特徴語が得られるか確認する．また，SVM によって語が重要かどうか推定し，正解と不正解の場合で指標の傾向の違いを確認する．分析には，男性が市道で男に刺され死亡したという通り魔事件の記事を用いる．

5.1 指標の主成分分析

評価実験の入力文章に用いたリード 20 記事のうちの 1 つに対して評価実験と同様に DF，LDA，格フレーム (CFC) を用いたモデルの指標を求め，主成分分析を行い，各モデルの傾向を比較する．第一主成分を横軸，第二主成分を縦軸として語をプロットし，指標ごとの因子負荷量を矢印で図 3 に示す．

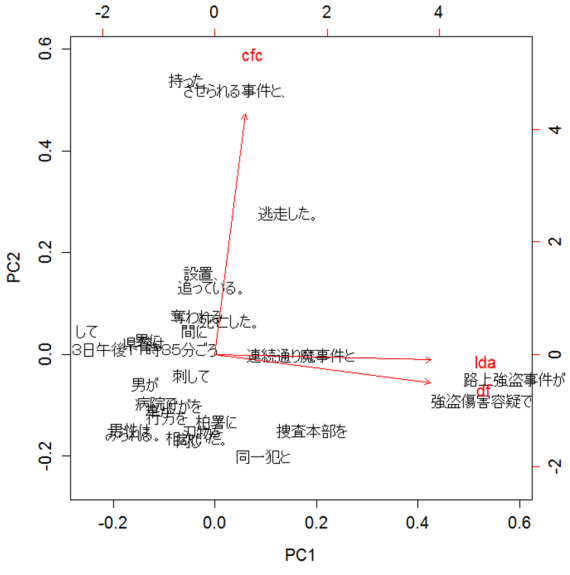

図 3　主成分分析

図 3 より，格フレームモデルの因子負荷量のベクトルは他のベクトルと垂直となっている．これより格フレームモデルの傾向は他と独立していることが分かる．また，DF，LDA モデルの因子負荷量のベクトルを見ると，互いのベクトルが水平であり，傾向が類似していることが分かる．DF，LDA モデルの傾向が類似しているのは，どちらもある条件における語の出現確率を指標としており，条件の部分が互いに独立でないためであると考える．

5.2 各モデルの特徴語

評価実験の入力文章に用いたリード 20 記事のうちの 1 つに対して評価実験と同様に DF，LDA，格フレームを用いたモデルの指標を求め，各モデルによる指標の上位に出現する形態素に注目して各モデルの特徴を掴む．表 4〜6 に各モデルの形態素ごとの指標の平均値を求め，上位 20 件を示す．

表 4　DF モデルの特徴語

形態素	$Score_{DF}$
する	0.00223
午後	0.00120
持つ	0.00098
みる	0.00089
男性	0.00073
同じだ	0.00060
車	0.00055
設置	0.00044
男	0.00042
病院	0.00040
県警	0.00040
間	0.00038
数詞	0.00036
相次ぐ	0.00031
追う	0.00029
奪う	0.00025
けが	0.00020
行方	0.00017
殺人	0.00014
首	0.00013

表 5　LDA モデルの特徴語

形態素	$Score_{LDA}$
する	0.01844
魔	0.01346
連続	0.01346
男性	0.01125
県警	0.00974
署	0.00845
男	0.00737
みる	0.00703
事件	0.00481
柏	0.00461
殺人	0.00344
午後	0.00335
病院	0.00316
行方	0.00238
車	0.00225
持つ	0.00148
犯	0.00145
同一だ	0.00145
数詞	0.00108
刺す	0.00108

表 4 より DF モデルの指標は「男性」「車」など一般的な語に高い点を与えている．ただし，「する」などの特別に意味の無い不要語にも高い点を与えている．

表 5 より LDA モデルの特徴語は DF モデルの特徴語と比較して「県警」「殺人」など事件性のあるトピックに関連する語が上位に来ている．また，トピック内での出現確率であるため，LDA モデルの上位の特徴語は全体的に DF モデルよりも指標が高いことが分かる．

表 6 より格フレームモデルの指標は「刃物」「行方」など目的格になるような述語に対して関連性の強い語の指標が高いことが分かる．これらより，各モデルが提案手法で意図したような特徴を持つ語を高く評価していることが分かる．

5.3 正解語の傾向

評価実験の入力文章に用いたリード 20 記事に対して評価実験と同様に DF，LDA，格フレームを用いたモデルの指標を求め，リード 19 記事についてモデルの指標を特徴量として，データセットを作成し，グリッドサーチによりパラメータを決定する．決定したパラメータを用いて SVM で学習を行い，残りの

表 6　格フレームモデルの特徴語

形態素	$Score_{CFC}$
同じだ	0.35863
みる	0.29099
犯	0.29099
同一だ	0.29099
相次ぐ	0.28341
刃物	0.16478
男性	0.10824
行方	0.08153
本部	0.06882
捜査	0.06882
車	0.06213
首	0.05003
署	0.04588
けが	0.03964
病院	0.03808
柏	0.02294
刺す	0.00541
男	0.00441
容疑	0.00291
傷害	0.00291

リード 1 記事の解析を行う．語が重要かどうかの推定より，正解と不正解の場合で指標の傾向の違いを確認する．DF モデルの指標が 10^{-4} 以上の場合について，正解と不正解を系列とし，LDA，格フレームの指標に $-log_{10}$ をとった値を横軸，縦軸とした散布図を図 4 に示す．

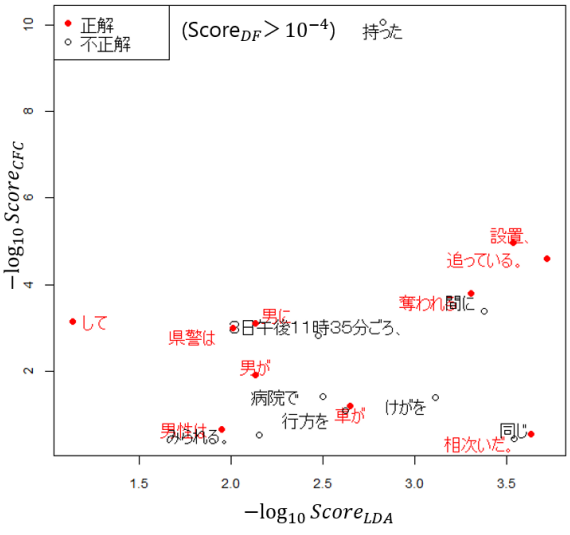

図 4　指標の傾向の散布図

図 4 より DF モデルの指標が大きい場合，LDA モデルの指標が大きい ($-log_{10}(Score_{LDA})$ が小さい) とき正解となりやすく，また格フレームモデルの指標が小さい ($-log_{10}(Score_{CFC})$ が大きい) とき正解となりやすいことが分かる．格フレームモデルの指標が小さい語には述語が多く見られることより，頻出する語ではトピックへの関連性が高い，もしくは述語である語が重要と判断される傾向があることが考えられる．また散布図より特徴的な正解の LDA モデルの指標が最も大きい 3 語，格フレームモデルの指標が最も小さい 3 語を示す．

表 7　LDA モデルの指標が大きい正解語

	DF	LDA	CFC
男性は	0.00073	0.01125	0.21648
して	0.00891	0.07218	0.00069
県警は	0.00040	0.00974	0.00099

表 8　格フレームモデルの指標が小さい正解語

	DF	LDA	CFC
設置	0.00044	0.00029	0.00001
追っている	0.00029	0.00019	0.00003
間に	0.00038	0.00042	0.00040

表 7 より「して」をのぞいた語は事件性のあるトピックに関係していると考えられ，重要とされやすいことがわかる．表 8 ではトピック性が薄く，述語との関連性が低い語が正解として選ばれていることが分かる．

6. おわりに

DF，LDA，格フレームを用いた 3 つの統計モデルを特徴量とし SVM によって重要かどうかを推定する手法を提案した．TFIDF を用いるモデルをベースラインと比較した結果，提案手法ではベースラインの手法より高い性能で重要かどうかの推定が可能なことが確認できた．また，モデル単体の性能は低いが，組み合わせることで性能が高まることが明らかになった．具体的に新聞記事に対して各モデルで評価値を求め，各モデルの特徴を確認した．また，指標の傾向を確認したところ，DF と LDA モデルの指標の傾向が類似していることが分かった．モデル単体を適切な手法で正規化することが課題であると考える．

文　献

[1] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. Machine Learning, Vol. 20, No. 3, pp. 273-297, 1995.

[2] Zechner, K., Fast Generation of Abstrants from General Domain Text Corpora by Extracting Relevant Sentences In Proceedings of the 16th International Conference on Computational Linguistics, pp.986-989, 1996.

[3] 堀之内寛，山本幹雄，n-gram モデルと IDF を利用した統計的日本語文短縮，言語処理学会　第 6 回年次大会発表論文集，2000.

[4] 畑山満美子，松尾義博，白井諭，重要語句抽出による新聞記事自動要約，情報処理学会研究報告自然言語処理 (NL)，pp.95-101，2001.

[5] David M. Blei, Andrew Y. Ng and Michael I.Jordan, Latent dirichlet allocation, the Journal of machine Learning research, vol.3. pp.993-1022, 2003.