技術関連記事の自動要約

平前　歩† 　乗重　雅誉‡ 　難波　英嗣† 　竹澤　寿幸†

†広島市立大学大学院 情報科学研究科 〒731-3194 広島県広島市安佐南区大塚東 3-4-1
‡広島市立大学 情報科学部 〒731-3194 広島県広島市安佐南区大塚東 3-4-1
E-mail: {hiramae, norishige, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

あらまし　本研究では，技術関連のニュース記事に対してその要約を付与することで，ユーザーにとって重要な情報を容易に把握できるように支援するシステムを開発する．ある技術分野でどのような投資，買収，提携などが行われているのか，あるいはどのような技術開発が行われているのかという情報は，企業の経営者が経営判断をする際に重要になると言える．そこで本研究では，企業の経営者を支援するため，技術関連のニュース記事の種類を自動で判別し，種類に応じた要約手法を適用することで，より適切な要約の作成を試みる．

キーワード　ニュース記事要約，機械学習，情報抽出，文書分類

1. はじめに

企業の経営者にとって，他の関連企業がどのような技術を開発しているのか，あるいは関連企業の買収，投資，技術提携などの状況を把握しておくことは，経営判断をする上で重要である．このような情報を入手するための情報源の一つとして新聞記事がしばしば使われるが，膨大な技術関連記事の中から自分の興味のある分野のものを見つけ，その全てに目を通すのは労力がかかる．そこで本研究では，効率的な情報収集を実現するためのシステムを構築する．

本システムではまず，技術関連記事を先行研究の技術[1]を用いて，各記事に国際特許分類を付与する．次に，「投資」，「提携」，「買収」，「技術」の 4 つのカテゴリに自動分類する．各カテゴリに関する説明を表 1 に示す．さらに，例えば「投資」に関する記事であれば「組織名（投資元）」，「投資金額」，「投資対象」といった情報を記事から自動抽出し，それらを表としてまとめ出力する．また，各カテゴリに応じて異なる要約手法を適用することで，技術関連記事の要約を作成し，出力する．このシステムを用いることで，ある分野の投資，提携，買収，技術に関する情報を表や要約で概観できるため，技術動向の把握が容易になると考えられる．

本論文の構成は以下の通りである．2 章では本研究で構築したシステムの動作例について，3 章では関連研究について，4 章では技術関連記事の自動要約について述べる．5 章で評価実験について述べ，6 章で実験結果について議論し，7 章で本論文をまとめる．

2. システム動作例

本章では，実装したシステムの概要と動作例について説明する．図 1 は「投資」に関する記事一覧，図 2 は「買収」に関する記事一覧をそれぞれ表示した例である．図 1 の例では，1 列目に記事の日付，2 列目に投資元の企業名，3 列目に投資先，4 列目に記事の見出しが表示される．例えば，1 行目の記事は 2006 年 1 月 12 日の記事を示しており，シャープが液晶事業に対して 5000 億円以上を投資したと読み取ることができる．このように表形式で表示することにより，素早い情報把握を支援することができる．また，記事の見出しが記事本文へのリンクになっており，より詳細な情報を知りたい時にクリックすることで，記事の全文を読むことができる．図 1，図 2 のようなシステムを「投資」，「提携」，「買収」，「技術」それぞれについて構築しており，各カテゴリに応じた情報を可視化するようになっている．

表 1：本システムで定義したカテゴリ

カテゴリ	説明
投資	企業，大学への投資，出資
提携	企業間，企業と大学間などの業務・企業提携
買収	企業買収
技術	新しく開発した技術，製品
その他	上記以外，一般論（論説，ブログ等）

日付	企業	投資先	金額	見出し
2006/01/12	シャープ	液晶事業	5000億円以上	薄型テレビ:大手家電メーカー、覇権争い 大型設備投資、活発に
2006/01/28	日立製作所	薄型テレビ用のプラズマパネル プラズマパネル	1000億円	日立製作所:プラズマパネル、生産増強へ新工場 1000億円規模投資
2006/02/10	東芝	半導体事業 半導体	630億円	東芝:半導体に630億円追加投資 一年間過去最大更新

図 1：システムの動作例（投資）

日付	買収元	買収先	金額	見出し
1993/01/14	ナブコ	ランソン・インダストリーズ社 グループニ社 米社	約十億円	ナブコ、米社を買収
1993/01/21	新日鉄	エヌ・エム・ビーセミコンダクター ミネベアの子会社	百数十億円	新日鉄が半導体事業に本格進出 ミネベアの子会社を近く買収
1993/01/23	伊藤忠商事 コロネット商会 伊藤忠 コロネット	ミラ・ショーン社 ミラ・ショーン	数億円	伊藤忠、ミラ・ショーン買収 コロネットと共同出資 消費低迷で救済

図 2：システムの動作例（買収）

3. 関連研究

3.1. 新聞記事に対する固有表現抽出と文書分類

固有表現抽出や文書分類は，古くから研究されている自然言語処理技術であり，様々なテキストを対象とした研究が多く存在する．ここでは，新聞記事を対象とした関連研究を紹介する．

これまでに，新聞記事から社会現象や話題語の抽出に関する研究は盛んに行われている[2][3][4]．これらの研究では，新聞記事集合のクラスタリング結果を tf*idf 解析し，最新話題語や課題発見を行っている．新聞記事の分類という点では本研究と類似しているが，本研究では一つのニュース記事を処理の対象としており，機械学習を用いて分類を行う．また，組織名や投資，買収にかかる金額などを，機械学習手法を用いて抽出している点で異なる．

斎藤ら[5]は，新聞記事集合から「祭り」,「コンサート」のようなイベントについて書かれた記事を自動で判定し，その記事からイベント名，開催日時などの情報を抽出する手法を提案している．イベント記事の判定や，記事からのイベント情報の抽出には手掛かり語を用いた機械学習手法を用いている．本研究においても，手掛かり語を用いた機械学習によって技術関連記事から組織名などの重要表現の抽出を行う点で関連している．しかし，斎藤らはイベント記事かそうでないかの 2 値分類を行っているが，本研究では技術関連記事かどうかの 2 値分類を行うのではなく，技術分野に関する 4 カテゴリと,「その他」の 1 カテゴリを合わせ，5 値分類を行う点で異なる．

3.2. テキスト要約

テキスト要約は，テキストの本文から重要な情報のみを抽出し,要点の素早い把握を支援する技術である．TAC[1], TSC[6][7]のようなテキスト要約を共通課題とした評価型ワークショップも開催されており，活発に研究がなされている分野である．本節では，代表的なテキスト要約手法について説明する．

[1] http://www.nist.gov/tac/

LexRank

Erkan ら[8]は，グラフベースの重要文抽出手法を提案している．重要文抽出手法は，テキスト中から重要な文や段落を抜き出し，適切な順に並べて出力したものを要約とする手法である．LexRank はまず，対象テキストに含まれる文間の類似度を計算し,文をノード,文間の関係をエッジとした類似度グラフを作成する．次に，類似度が閾値以上であれば 1，それ以外は 0 を要素とする隣接行列を用意する.作成したグラフから,隣接行列に対してべき乗法を用いて主固有ベクトルを計算することで,ノードの重要度を得る.LexRank は,単に字数の多いノードを評価するだけでなく，字数の多いノードと隣接しているノードの重要度も考慮している．つまり，LexRank によって計算される文の重要度は，他の多くの文と類似する文ほど高く，さらに重要度の高い文と類似する文の重要度も高くなる．本研究の提案手法は，この LexRank をベースとしている．

ILP（Integer Linear Programming）

Woodsend ら[9]は，テキスト要約を最適化問題の一つである整数計画問題に定式化している．ニュース記事中の重要文を選択して要約に含めるのではなく，フレーズ単位で選択する要約手法を提案している．対象となるニュース記事に含まれるフレーズの重要度をあらかじめ算出しておき，その重要度が制約条件を満たす中で最大となるようにフレーズを選択することで，要約を作成している．

Deep Learning

近年，深層学習を用いたテキスト要約に関する研究が見られる．深層学習は機械翻訳で一定の成果が得られて以降，画像や対話応答など多くの系列生成タスクにおいて用いられている．Rush ら[10]が Annotated English Gigaword Corpus に含まれるニュース記事から,大量の訓練事例の自動構築を行ったことにより，テキスト要約についても encoder-decoder モデルのような深層学習に基づいた研究が増えている．Nallapati ら[11]は，要約対象のテキストに含まれる重要単語を考慮する場合や複数の文を入力する場合など，様々なパターンを用いて幅広い調査を行った．また，Nallapati らはこの研究を発展させて，複数テキスト要約を対象とした研究[12]や，Recurrent Neural Network (RNN) に基づいた sequence-to-sequence モデルを用いた要約システムである SummaRuNNer[13]を構築している．
他にも，要約対象テキストの内容を元に新たな文を生成して要約を作成する，生成型要約に関する研究も行われている．Rush ら[10]は Attention 機構を導入した encoder-decoder モデルを用いた文レベルの生成型要約手法を，吉岡ら[14]は複数テキストをから一文要約の

生成を行う手法をそれぞれ提案している．

深層学習に基づいたテキスト要約における問題点として，出力する要約の長さを制御できないことが挙げられる．現状では，要約モデルの訓練に用いた原文と要約の長さに依存してしまうため，ユーザーによって入力された要約長に合わせた要約を生成することが難しい．菊池ら[15]はこの問題を解決するために，encoder-decoder モデルに出力長制御の機能を，学習によって獲得させる手法を提案している．これまでに挙げた深層学習に基づく要約手法は，英語のテキストを対象としている．本研究における提案手法は日本語で書かれたテキストを対象としており，Deep Learning ではなく，従来のテキスト要約手法をベースとしている．

4. 技術関連記事の自動要約

本研究で開発するシステムは，技術関連記事を表 1 の 5 カテゴリに分類するモジュール，各記事から組織名などを抽出するモジュール，各記事の要約を作成する要約モジュールから構成される．本章では，4.1 節でシステム概要，4.2 節で技術関連記事に対する固有表現抽出，4.3 節で技術関連記事のカテゴリ分類，4.4 節で技術関連記事の要約作成について説明する．

4.1. システム概要

システム全体の処理手順として，まず，固有表現抽出を行い，その結果を用いて記事をカテゴリに分類する手法（NE→Cls 法）と，先に記事をカテゴリ分類した後に，カテゴリに応じた固有表現抽出を行う手法（Cls→NE 法）が考えられる．なお，要約作成は NE→Cls 法，Cls→NE 法のいずれにおいても，システムの最後の手順とする．表 2 に NE→Cls 法，Cls→NE 法を実現する上で用いる手法を示す．各手法の説明は次節以降で行う．

本研究で構築するシステムについて，図 3 を用いて説明する．図 3 は Cls→NE 法を適用した後に要約作成を行うシステムの概要であり，入力のニュース記事が分類モジュールによって「買収」に分類され，「買収」用の固有表現抽出モジュールと要約モジュールによって要約が出力される場合の例を示している．本研究では技術関連記事を対象とするため，その他に分類された記事は要約を作成しない．

表 2：NE→Cls 法，Cls→NE 法を実現する上で用いる手法

手法	分類モジュール	固有表現抽出モジュール
Cls→NE 法	fastText[16]	CRF
NE→Cls 法	ルールベース	CRF

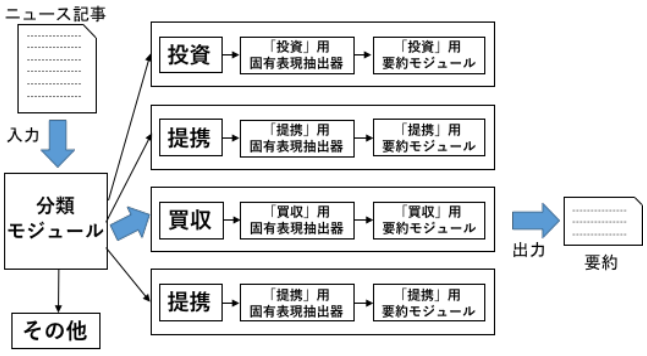

図 3：本研究で構築したシステムの概要（Cls→NE 法）

4.2. 技術関連記事に対する固有表現抽出

本研究では，技術関連記事に含まれる重要単語に対してタグを自動付与する固有表現抽出器を構築する．まず，抽出すべき固有表現を定義するために，人手による技術関連記事の分析を行った．具体的には，表 1 で定義したカテゴリに分類し，「その他」以外に分類された技術関連記事について，記事に含まれる重要単語に対してタグ付けを行った．「投資」，「買収」，「提携」，「技術」それぞれの記事の例を図 4，図 5，図 6，図 7 に示す．

（2006/1/12 読売新聞）

薄型テレビ：大手家電メーカー、覇権争い　大型設備投資、活発に

デジタル家電の需要急増を受けて、大手電機メーカーの間で大型設備投資に踏み切る動きが活発になってきた。特に松下電器産業はプラズマ、シャープは液晶に集中投資し、世界の薄型テレビの覇権を争う構えだ。◆液晶で勝負<組織名(投資元)>シャープ</組織名(投資元)>の町田勝彦社長は１１日、０８年度までの３年間で<投資対象>液晶事業</投資対象>だけで<投資金額>５０００億円以上</投資金額>を投資し、…

図 4：「投資」に分類された記事にタグを付与した例

（1993/1/14 読売新聞）

<組織名(買収元)>ナブコ</組織名(買収元)>、<組織名(買収先)>米社</組織名(買収先)>を買収

<組織名(買収元)>ナブコ</組織名(買収元)>は十三日、米国の中堅自動ドアメーカーの<組織名(買収先)>ランソン・インダストリーズ社</組織名(買収先)>（本社・ウィスコンシン州）と<組織名(買収先)>グループ二社</組織名(買収先)>を、<買収金額>約十億円</買収金額>で買収したと発表した。北米に約二百か所の拠点を持つランソン社の販売力を活用して、事業拡大を図るのが狙い。

図 5：「買収」に分類された記事にタグを付与した例

（2006/1/13 読売新聞）
<組織名>楽天</組織名>：損保提携先は<組織名>ＡＩＧ</組織名> 総合ネット金融、態勢整う
損害保険事業への参入を目指している<組織名>楽天</組織名>の提携先が、米国の保険グループ<組織名>ＡＩＧ</組織名>に固まったことが１２日、明らかになった。…

図 6:「提携」に分類された記事にタグを付与した例

（1993/1/4 読売新聞）
風疹 胎児感染に遺伝子診断 <組織名>国立予防研</組織名>が開発 過去に２０人以上出産
妊娠初期に風疹（ふうしん）の母子感染の有無がわかる遺伝子診断法の開発に、<組織名>国立予防衛生研究所村山分室</組織名>（東京都武蔵村山市）のグループがわが国で初めて成功した。…

図 7:「技術」に分類された記事にタグを付与した例

技術関連記事の分析結果より，本研究で抽出すべき固有表現は表 3 のように定義した．本研究では，技術関連記事における重要単語を抽出する課題を固有表現抽出問題とみなし，機械学習手法の一つである CRF（Conditional Random Fields）を用いる．CRF に与える素性は，以下の通りである．

- ターゲットとなる形態素から，前後 3 形態素のユニグラム，バイグラム，トライグラム
- ターゲットとなる形態素から，前後 3 形態素の品詞
- 日本語係り受け解析器 CaboCha の固有表現抽出機能によって地名（LOCATION）タグ，組織名（ORGANIZATION）タグが付与された形態素

表 3：本研究で抽出する固有表現

カテゴリ	固有表現
投資	投資金額 組織名（投資元） 投資対象
提携	提携する組織名
買収	買収金額 組織名（買収元） 組織名（買収先）
技術	組織名（開発元）

4.3. 技術関連記事のカテゴリ分類

本研究では，技術関連記事を表 1 の 5 カテゴリに分類し，一つのニュース記事は一つのカテゴリにだけ分類されるようにする．本節では，Cls→NE 法，NE→Cls 法それぞれの分類モジュールについて説明する．

Cls→NE 法における分類モジュールには， Joulin らが提案している fastText を用いる．fastText は入力層，隠れ層，出力層の 3 層からなる DNN(Deep Neural Network)をベースとしており，単語の分散表現の学習やテキストをあらかじめ決めたカテゴリに分類することができる．Piotr ら[17]の研究では，fastText は Word2Vec とその類型モデルでそれまで考慮されていなかった,「活用形」まとめられるようなモデルになっている．例えば，go，goes，going は全て go の活用形だが，字面的にはすべて異なるのでこれまでの手法では別々の単語として扱われてしまう．そこで，単語を構成要素に分割したものを考慮することで，字面の近しい単語同士により意味のまとまりをもたせるという手法を提案している．また，日本語テキストに対して用いる場合は分かち書きして半角スペース区切りにしておく必要があり，それぞれの文章がどのクラスにあるかという教師データが必要となる.カテゴリ分類は，一つだけでなく複数のカテゴリに分類することも可能である.

次に，NE→Cls 法において，カテゴリ分類を行う上でのルールを説明する．なお，ルールは記事のタイトルと本文両方に対して適用することでカテゴリを判定する．ルールは以下の通りである．

①：表 3 で定義した固有表現タグの有無を確認する．あれば②を適用する．無ければ「その他」とする．

②：記事における金額タグの有無を確認する．あれば③，無ければ④を適用する．

③：投資対象タグの有無を確認する．あれば「投資」に分類し，無ければ「買収」とする．

④:提携という単語の有無を確認する.あれば「提携」，無ければ「技術」とする．

NE→Cls 法は 4.1 節で述べた通り，固有表現抽出の結果を用いるが,「提携」と「技術」はどちらも組織名のみ定義しているため，固有表現抽出結果だけでは判定ができない.そのため,④のみ手がかり語を利用した．

4.4. 技術関連記事の要約作成

本研究で構築するシステムにおいて抽出すべき情報は，4.2 節で示した固有表現抽出による結果だけでは不十分な場合がある．例えば，投資や買収などを行った理由，狙いが挙げられる．このような例は，単語だけで表現することは難しく，文を読んで初めて理解ができると考えられる．固有表現抽出はテキストに含まれる単語を抽出する技術のため，文単位で抽出することができない．このような理由から，4.2 節，4.3 節で説明した手法により「投資」,「提携」,「買収」,「技術」のいずれかに分類された技術関連記事を対象に，それぞれの技術関連記事の要約を作成する．本研究では，3.2 節で説明した LexRank をベースとした手法を

提案する．LexRank は文の類似度グラフを用意して，ノードの PageRank 値を計算することで文の重要度を算出する．文の類似度グラフを作成する際，文間の類似度は，文を tf*idf 値を要素とするベクトルで表し，類似性尺度にコサイン類似度を用いて計算する．

例えば，「技術」に分類された記事では，技術開発を行った組織について言及している文の重要度を高く，「投資」に分類された記事では投資元の組織名，投資対象，金額に関する文の重要度を高く設定すべきだと考えられる．そのために，まず，4.3節で説明した固有表現抽出手法により技術関連記事に対してタグを付与する．この時タグが付与された文は重要な情報を含んでいると言えるため，このような文の重要度を高く設定する．このように，技術関連記事のカテゴリと文の重要度を考慮するために，Biased LexRank[18]を用いる．まず，タグが付与された文から，タグの個数をカウントする．そして，PageRank値を計算する際，タグの個数を元に各ノード（文）へのランダムジャンプ確率に設定する．これにより，タグを含む文を示すノードへのランダムジャンプ確率を上げ，重要度を高くすることができる．

5. 評価実験

本章では，本研究で行った評価実験について説明する．5.1 節で固有表現抽出モジュールに関する実験，5.2 節で分類モジュールに関する実験，5.3 節で NE→Cls 法，Cls→NE 法に関する実験，5.4 節で要約作成に関する実験についてそれぞれ説明する．

5.1. 技術関連記事に対する固有表現抽出

毎日新聞，日本経済新聞，読売新聞から人手でニュース記事を収集し，「投資」，「提携」，「買収」，「技術」に分類した．次に，収集した技術関連記事に対して，表 3 で示した固有表現タグを人手で付与した．表 4 に本実験に使用する技術関連記事の件数，技術関連記事に対して付与したタグの件数を示す．機械学習には CRF を用い，2 分割交差検定を行った．また，評価尺度には精度，再現率を用いた．

表 4：5.1 節の実験に用いた技術関連記事データ

	記事件数	組織名タグ数	買収先タグ数	投資対象タグ数	金額タグ数
投資	110	218	-	57	175
提携	104	412	-	-	-
買収	203	290	278	-	92
技術	155	353	-	-	-

5.2. 技術関連記事のカテゴリ分類

表 4 の技術関連記事データに，人手で収集した「その他」に適切なニュース記事を 128 件追加したデータを実験に用いる．本実験のデータは，固有表現タグは付いていない状態で使用した．本実験では，4 章で述べた NE→Cls 法と Cls→NE 法を実施し，カテゴリ分類に関する実験を行った．本実験はカテゴリ分類に関しての評価を行うため，NE→Cls 法において固有表現抽出は 10 割の精度，再現率の再現率が出ているものと仮定する．5 分割交差検定を行い，評価尺度には精度，再現率を用いた．

5.3. カテゴリ分類と固有表現抽出

5.1 節，5.2 節で行った実験に加え，新たに用意したテストデータで Cls→NE 法，NE→Cls 法を再度適用して実験を行った．テストデータはオンラインニュースサイトから記事を人手で収集し，技術関連記事の選別を行い，固有表現タグを付与した．表 5 に本実験に使用する技術関連記事の件数を示す．本実験は 5.1 節，5.2 節の実験を合わせており，最終的な評価は付与された固有表現タグを基準に行う．例えば，記事のカテゴリが正解と違っていた場合，タグの付与が合っていても不正解とみなす．つまり，記事のカテゴリが正解している場合のみ，付与されたタグが正解かどうか判定する．評価尺度は精度，再現率を用いる．

表 5：5.3 節の実験に用いた
技術関連記事データ

	記事件数
投資	3
提携	21
買収	17
技術	9
その他	69
合計	119

5.4. 技術関連記事の要約作成

表 4 で示した技術関連記事データから「その他」以外の 4 カテゴリについてそれぞれ 20 件選択した．選択した技術関連記事について，100 字以内の要約を人手で作成し，ROUGE[19]による評価を行うことで提案する要約手法の有効性を検証した．本実験のデータにおいても，固有表現タグは付いていない状態で使用した．実験データに用いる記事の文字数の平均は 517 文字，正解要約の文字数の平均は 94.5 文字だった．そのため，平均の要約率は 0.18 となった．ROUGE-N は，正解要約とシステム出力結果の両方に共通して含まれる N-gram の数を，正解要約中の N-gram の数で割った値を指す．比較手法を以下に示す．いずれの手法でも，要約長制限は 100 文字に統一した．なお，LexRank で文の重要度を計算する際のダンピングファクタは 0.85 とした．

- Lead 法（baseline）：技術関連記事の先行から文を抜き出す手法．
- LexRank（baseline）：単純に LexRank を適用し，文の重要度を計算する手法．
- ILP（baseline）：文の要約を整数計画問題ととらえ，制約条件の中で文の重要度が最大となるように文を選択する手法．
- LexRank+NE：4.3 節で示した固有表現抽出手法により記事にタグを付与し，タグを利用した LexRank により，文の重要度を計算する手法．

6. 実験結果と考察

6.1. 技術関連記事からの固有表現抽出

5.1 節で説明した実験の結果を表 6 に示す．表 6 を見ると，組織名，金額はどのカテゴリにおいても 0.7 ポイント前後の精度，再現率を得られている．「投資対象」において再現率が 0.146 ポイントという結果になった．この実験結果を分析したところ，「半導体事業」のように名詞のみで構成されている固有表現は抽出できたが，「IT システムのための研究・開発」のように名詞以外の品詞が含まれる固有表現は抽出できていなかった．このような固有表現の抽出は，CRF をベースとした抽出器を用いるのではなく，他の手法を用いた抽出器を構築することで解決を図る必要がある．

表 6：固有表現抽出モジュールの実験結果

カテゴリ	タグ	精度	再現率
投資	金額	0.780	0.821
	組織名（投資元）	0.884	0.702
	投資対象	0.609	0.146
提携	組織名	0.891	0.773
買収	金額	0.648	0.686
	組織名（買収元）	0.728	0.597
	組織名（買収先）	0.565	0.310
技術	組織名	0.845	0.752

6.2. 技術関連記事のカテゴリ分類

5.2 節で説明した実験の結果を表 7 に示す．表 7 から分かる通り，NE→Cls 法では，精度，再現率共に 0.807，Cls→NE 法では精度，再現率共に 0.833 を得ることができた．この結果より，本システム全体の処理手順として，Cls→NE 法の有効性を確認した．NE→Cls 法について，検出誤り例を示して考察を行う．図 8 は，本来「投資」に分類されるべき記事であるが，「技術」に分類された記事を示している．図 8 の例では，金額タグが付与されていないため，4.3 節で説明したルール②が適用されなかったことが原因である．本研究で設定したルールでは，最初に金額タグの有無によって「投資」か「買収」，「技術」か「提携」の 2 パターンに分けるため，金額タグの付き方に大きく左右されてしまう．また，本実験では NE→Cls 法において固有表現抽出が 10 割の精度，再現率が出ている前提で行ったため，カテゴリ分類において 8 割程度の精度，再現率を得ることができたが，実際に固有表現抽出モジュールを適用した際は，固有表現抽出の精度が下がるため，カテゴリ分類の精度も低下すると考えられる．このような例は，Cls→NE 法のように先に技術関連記事のカテゴリ分類を行い，カテゴリに応じた固有表現抽出手法を用いることで解決を図れるため，Cls→NE 法が NE→Cls 法より良い結果を示したと考えられる．

表 7：カテゴリ分類モジュールの実験結果

	精度	再現率
NE→Cls 法	0.807	0.807
Cls→NE 法	0.833	0.833

（2001/6/6　読売新聞）
<組織名>松竹</組織名>、韓国映画「春の日は過ぎゆく」に投資　【ソウル共同】
日本の映画製作大手、<組織名>松竹</組織名>（大谷信義社長）は５日、韓国の映画製作会社「<投資対象>サイダス</投資対象>」の新作「春の日は過ぎゆく」に、香港の企画会社「<組織名>アプローズピクチャーズ</組織名>」とともに投資することで合意、ソウル市内のホテルで調印式を行った。　…

図 8：検出誤り例（正解：「投資」，実験結果：「技術」）

6.3. カテゴリ分類と固有表現抽出

5.3 節で説明した実験の結果を表 8 に示す．今回の実験においては，カテゴリ分類実験の結果が正解と一致しなかった場合はタグの付き方が仮に正解だとしても，不正解としてカウントしている．また，「その他」が正解の記事が，実験結果では誤分類されたため、誤ったタグが Cls→NE 法では 48 件，NE→Cls 法では 218 件付与されていた．NE→Cls 法では「技術」が正解の一部の記事が，誤分類されたため，4 件の組織名(買収元)タグ，1 件の投資対象タグが誤って付与されていた．「投資」の記事は，カテゴリ分類が全て誤っていたため，精度，再現率ともに 0 となった．

タグの付き方に着目すると，正解が「<組織名>エイチ・ツー・オー（Ｈ２Ｏ）リテイリング</組織名>」に対し，実験結果が「<組織名>エイチ・ツー・オー</組織名>」のように，部分的にタグが付与される例が存在した．上記の例であれば，「エイチ・ツー・オー」だけで組織名が一意に定まると言えるため，組織名の後ろのかっこで囲まれている文字列は省いても良いと考

えられる．そのため，組織名だけにタグが付くよう統一する必要がある．

Cls→NE 法の実験において，NE→Cls 法より精度が高く，再現率は低いという結果が得られた．しかし，タグ別に結果を見ると，「提携」の組織名以外は Cls→NE 法が良い結果を示した．結果より，カテゴリに特有の固有表現抽出モジュールが有効であると言える．ここで，「提携」の組織名の結果について考察を行う．Cls→NE 法では，「提携」に分類されるべき記事が「その他」，「技術」に誤分類された例が 21 件中 13 件存在した，これにより不正解となるタグが増えたことが原因と考えられる．

NE→Cls 法の実験においては，「その他」に分類された記事が 0 件だった．固有表現タグが 1 つもつかないことが「その他」に分類される条件であるため，全ての記事に何らかのタグが付いていることがわかる．実際に，全体の正解タグが 201 件に対して，NE→Cls 法による実験結果で付与されたタグは 469 件存在した．そのため，精度が低く，再現率が高い結果となったと考えられる．本研究で構築するシステムは，技術関連記事かそうでないかの判断が必要になるため，「その他」への分類を適切に行える NE→Cls 法が適していると考えられる．

表 8：Cls→NE 法，NE→Cls 法を適用した結果

		Cls→NE 法		NE→Cls 法	
カテゴリ	タグ	精度	再現率	精度	再現率
投資	金額	0.000 (0/0)	0.000 (0/4)	0.000 (0/3)	0.000 (0/4)
	組織名	0.000 (0/0)	0.000 (0/5)	0.000 (0/8)	0.000 (0/5)
	投資対象	0.000 (0/0)	0.000 (0/4)	0.000 (0/0)	0.000 (0/4)
提携	組織名	0.364 (28/77)	0.315 (28/89)	0.508 (61/120)	0.685 (61/89)
買収	金額	0.350 (7/20)	0.438 (7/16)	0.292 (7/24)	0.438 (7/16)
	組織名 (買収元)	0.842 (16/19)	0.432 (16/37)	0.215 (14/65)	0.378 (14/37)
	組織名 (買収先)	0.700 (14/20)	0.424 (14/33)	0.538 (7/13)	0.212 (7/33)
技術	組織名	0.500 (7/14)	0.538 (7/13)	0.308 (4/13)	0.308 (4/13)
全体	-	0.364 (72/198)	0.358 (72/201)	0.198 (93/469)	0.463 (93/201)

6.4. 技術関連記事の要約作成

5.4 節で説明した実験の結果を表 9, 表 10 に示す．まず，全体の平均を見ると，ROUGE-1，ROUGE-2 ともに LexRank＋NE がベースラインである ILP, LexRank を上回ったが，Lead 法を上回ることはできなかった．しかし，LexRank+NE は固有表現抽出の結果を利用したことにより，Lead 文でない，カテゴリ特有の固有表現を含む重要文が選ばれやすくなったと考えられる．

カテゴリ別に ROUGE-1, ROUGE-2 の結果を見ると，「投資」，「買収」，「技術」では LexRank+NE が LexRank と ILP より高い値を示したが，「提携」においては LexRank+NE が LexRank をより低い値になった．ここで，「提携」の実験結果について考察を行う．実験データ 20 件のうち，ROUGE-1 による評価において，LexRank より LexRank+NE の方が，評価値が高かった場合は 5 件，同じ値の場合が 10 件，低かった場合が 5 件となった．ROUGE-2 では，評価値が高かった場合は 5 件，同じ値の場合が 11 件，低かった場合が 4 件となった．このように，全体平均としては評価値が下がっても，LexRank+NE によって良い結果が得られる場合が「提携」全体で 3 分の 1 存在することがわかるため，LexRank+NE が「提携」において大きく劣っているとは言えない．

全カテゴリについて LexRank と LexRank＋NE を適用した結果を比較したところ，LexRank+NE は固有表現抽出の結果によって適切に補正がかかったことにより，技術関連記事から適切に重要文を抽出できた例があった他，重要でない文から組織名や金額を抽出したことにより，間違った補正がかかった例が存在した．他にも，間違った固有表現抽出が行われると，本来重要度が低くなっているべき文の重要度が上がってしまうため，間違った要約が作成されてしまう．「提携」に関する記事ではこの傾向が特に多く見られた．このように，固有表現抽出器によって付与されるタグの付き方によって結果が左右されるため，固有表現抽出の精度，再現率を上げることで改善ができると考えられる．

今回の実験では，100文字に設定したことによって選ばれる文が1文だけの例が多く存在した．4.4節で説明した通り，企業動向を説明する文だけでなく，投資した理由のように動機を示す文が要約に含める必要があると考えられるが，今回の実験では，要約長を100文字に設定したことによって選ばれる文が1文だけの例が多く存在した．そのため，企業動向を説明する文しか要約に含めることができなかった．今後は，適切な要約長を検討する他，動機を示す文を自動抽出する手法を検討する必要がある．

表 9：技術関連記事の要約作成実験結果（ROUGE-1）

要約手法	投資	提携	買収	技術	平均
Lead 法 (baseline)	0.597	0.612	0.486	0.578	0.568
ILP (baseline)	0.327	0.361	0.318	0.257	0.316
LexRank (baseline)	0.321	0.457	0.369	0.398	0.386
LexRank+NE	0.493	0.439	0.399	0.445	0.444

表 10：技術関連記事の要約作成実験結果（ROUGE-2）

要約手法	投資	提携	買収	技術	平均
Lead 法 (baseline)	0.482	0.473	0.342	0.430	0.419
ILP (baseline)	0.179	0.227	0.156	0.130	0.173
LexRank (baseline)	0.164	0.309	0.210	0.262	0.236
LexRank+NE	0.353	0.288	0.260	0.324	0.306

7. おわりに

本研究では，技術関連記事のカテゴリを判別し，カテゴリに応じた固有表現抽出と要約を行うシステムを構築した．fastText によるカテゴリ分類，CRF による固有表現抽出の順に行う Cls→NE 法と，グラフベースの文の重要度計算手法である LexRank を拡張し，カテゴリ別の固有表現抽出の結果を利用した LexRank+NE によってカテゴリ別に適切な要約を作成する手法を提案した．今後の課題として，本研究では「投資」,「提携」,「買収」,「技術」,「その他」の 5 カテゴリを定義したが,「経営方針」,「政策」といった他の技術関連の情報にも着目することで，よりニーズに合ったニュース記事の要約システムの構築ができると考えられる．

謝辞

本研究は Panasonic 株式会社の支援を受けて行われた．

参 考 文 献

[1] Iinuma, S., Fukuda, S., Nanba, H., and Takezawa, T. (2015). Evaluation of the Industrial and Social Impacts of Academic Research Using Patents and News Articles. The International Association for Computers & Information Science (ACIS) International Journal of Computer & Information Science, Vol.16, No.1, 12-21.

[2] 水落大史, 井上悦子, 吉廣卓哉, 村川猛彦, 中川優. (2010). 新聞記事集合に対する時系列のトピック抽出. DEIM Forum, 論文集, D6-3.

[3] 中村智浩、平野孝佳、平手勇宇、山名早人. (2008). 単独記事フィルタリングを用いた時系列ニュース記事分類法の提案. 電子情報通信学会技術研究報告．PRMU, パターン認識・メディア理解, 108(94), 59-64.

[4] 橋本泰一、村上浩司、乾孝司、内海和夫、石川正道. (2008). 文書クラスタリングによるトピック抽出および課題発見. 社会技術研究論文集, 5, 216-226.

[5] 斉藤隆太, 石野亜耶, 難波英嗣, 竹澤寿幸. (2012). 新聞記事と Web からのイベント情報の自動抽出. 第 5 回 Web とデータベースに関するフォーラム (WebDB Forum).

[6] Fukushima, T., and Okumura, M. (2001). Text Summarization Challenge Text Summarization Evaluation in Japan. In North American Association for Computational Linguistics (NAACL2001), Workshop on Automatic Summarization, 51-59.

[7] Okumura, M., Fukushima, T., and Nanba, H. (2003). Text Summarization Challenge 2: Text Summarization Evaluation at NTCIR Workshop 3. In Proceedings of the HLT-NAACL 03 on Text Summarization Workshop-Volume 5, 49-56, Association for Computational Linguistics.

[8] Erkan, G., and Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research, 22, 457-479.

[9] Woodsend, K., and Lapata, M. (2010). Automatic Generation of Story Highlights. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 565-574.

[10] Rush, A. M., Chopra, S., and Weston, J. (2015). A Neural Attention Model for Abstractive Sentence Summarization. Proceedings of EMNLP15, 379–389.

[11] Nallapati, R., Zhou, B., Santos, C. D., Gulcehre, C., and Xiang, B. (2016). Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. arXiv:1602.06023v2.

[12] Nallapati, R., Zhou, B., and Xiang, B. (2016). Sequence-to-Sequence RNNs for Text Summarization. International Conference on Learning Representations, Workshop Track.

[13] Nallapati, R., Zhai, F., and Zhou, B. (2016). SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. arXiv Preprint arXiv:1611.04230v1.

[14] 吉岡重紀, 山名早人. (2016). 生成型一文要約のためのマルチアテンションモデルの提案. DEIM Forum, 論文集, E8-3.

[15] 菊池悠太, 笹野遼平, 高村大也, 奥村学. (2016). Encoder-Decoder モデルにおける出力長制御. 研究報告自然言語処理 (NL), 2016(5), 1-9.

[16] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. arXiv Preprint arXiv:1607.01759.

[17] Piotr, B., Edouard, G., Armand, J., and Tomas, M. (2016). Enriching Word Vectors with Subword Information. arXiv:1607.04606v1 [cs.CL] 15.

[18] Otterbacher, J., Erkan, G., and Radev, D. R. (2009). Biased LexRank: Passage Retrieval Using Random Walks with Question-based Priors. Information Processing & Management, Vol. 45, No. 1, 42–54.

[19] Lin, C. Y. Lin. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. Proc. of Workshop on Text Summarization Branches Out, 74-81.