技術関連記事の分析に基づいた経営判断支援システムの構築

平前 歩　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

E-mail: {hiramae, 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の情報を表で概観できるシステムを提案する．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，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にはニュース記事や新聞記事がよく使われるが，膨大な数の技術関連記事の中から自分が興味のある情報を探し出し，その全てに目を通すのは労力がかかると考えられる．本研究では，深層学習を用いて技術関連記事を買収や投資などのカテゴリに分類した後，各カテゴリに応じた情報抽出を行い，抽出結果を表としてまとめて出力するシステムを構築した．

キーワード ニュース記事，深層学習，文書分類，情報抽出

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やオンラインニュース記事がしばしば使われるが，膨大な数の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時間と労力がかかる．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，ニュース記事を「投資」，「提携」，「買収」，「技術」，「その他」の 5 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」といった情報を技術関連記事から自動抽出し，それらを表としてまとめて出力する．表を見ることで，ある分野の企業動向を把握すること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 節では，本研究で構築したシステムの動作例を紹介し，3 節で関連研究について述べる．4 節では，本研究のシステムの構成と提案手法を説明し，5 節では提案手法の有効性を調べるために行った実験について述べる．6 節で結論，7 節で今後の課題について述べる．

2. システム動作例

本節では，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「買収」記事の一覧表，図 2 は「投資」記事の一覧表である．買収記事の一覧表では，1 列目に記事の日付，2 列目に買収元の企業名，3 列目に買収先の企業名，4 列目に買収金額，5 列目に記事の見出しが表示される．図 1 の例では，1 行目を見ることで，「ナブコがランソン・インダストリーズ社などを約十億円で買収した．」と読み取ることができる．システムは図 1，図 2 のような表をその他以外の 4 カテゴリで出力する．表を見ることにより，素早い情報把握が可能になるだけでなく，各企業が持つ技術の価値の判断，関連分野の企業動向の把握，経営判断をする上で役立つと考えられる．

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	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 グループ二社 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 1：システム動作例（買収）

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 2：システム動作例（投資）

3. 関連研究

本節では，深層学習を用いた文書分類，固有表現抽出に関する手法を紹介する．3.1 節で文書分類，3.2 節で固有表現抽出の手法をそれぞれ紹介する．

3.1. 文書分類

文書分類に関する研究では，古くからは SVM，ロジスティック回帰などの機械学習を用いた手法がよく利用されており，近年では深層学習を用いた手法が提案されている．以下より，深層学習を用いた文書分類手法を紹介する．

Kim[1]は，CNN(Convolutional Neural Network)をベースにし，評判分析や質問分類を行うモデルを提案している．提案モデルを図 3 に示し，動作を説明する．まず，文を単語に区切り，それらを単語埋め込み（Word Embedding）の列に置き換える． 次に，入力データを文の長さ n×単語埋め込みの次元数 k のベクトルでチャンネルを作成し，特徴マップを生成する．この時，ウインドウサイズをいくつか指定することで複数の特徴マップを生成している．その後，畳み込みを行い，Max-pooling で特徴マップの情報を圧縮して特徴量を獲得する．最後に Full connection 層において前の層の情報を全て結合し，softmax によって正規化することで最終的に分類先を決定する．Kim は Google News を word2vec[2]によって学習させたモデル（300 次元）を利用することで，性能が向上したと報告している．

Yang ら[3]は，評判分析に着目した階層 Attention ネットワークを提案している．Attention とは，LSTM や encoder-decoder モデルにおいて中間層の情報を大域的に取っておいて，必要なタイミングでそれらを利用する手法である．LSTM のような長期記憶が可能な手法では，解析が進むにつれて最初に解析した情報が減衰していくという問題点があるが，Attention を使うことでその問題点を解決することが可能になった．本研究においても深層学習に基づいた手法を用いるが，CNN や Attention は利用せず，3 層ニューラルネットワークに基づく手法を利用し，技術関連記事の分類を行う．

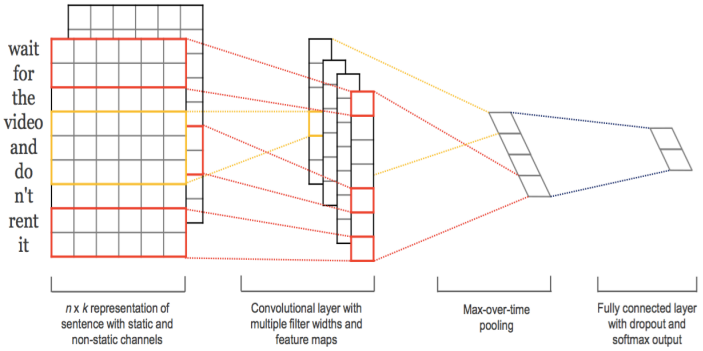

図 3：Kim らの提案モデル
(Kim et al.(2014)[1]より抜粋)

3.2. 固有表現抽出

固有表現抽出は，N 個の入力単語列$\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$から N 個の固有表現ラベルの列$\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$を予測する問題として定式化できる．このような問題は系列ラベリングと呼ばれる．系列ラベリングに関する研究の多くは CRF(Conditional Random Fields) [4]のような機械学習が用いられており，高い精度が得られることが知られている．

近年では，深層学習を用いた固有表現抽出を行うモデルが主流となっている．Lample ら[5]は，LSTM と CRF を組み合わせた LSTM-CRF を提案している．入力として単語埋め込みを LSTM に与え，LSTM の記憶セルなどで離れた距離の依存関係をモデル化することで，単語列に対して固有表現タグを付与する．また，Lample らは単に先頭単語から順方向に解析する LSTM だけでなく，最後の単語から逆方向に解析する LSTM を合わせた Bi-directional LSTM(以下 Bi-LSTM)を使用している．深層学習を用いた固有表現抽出においては，Bi-LSTM を使用することでモデルの性能が良くなることが知られている．

Ma ら[6]は，LSTM-CNNs-CRF というモデルを提案している．Ma らのモデルはまず，入力単語を構成する文字表現を CNN によって獲得する．次に，それらと単語埋め込みを合わせて入力とし，Bi-LSTM によって素性を獲得して CRF を適用する．このモデルは，英語のテキストを対象とした固有表現抽出モデルの中で最高性能を達成していることから，単語埋め込みと CNN で獲得した文字表現を組み合わせることの有効性を示している．

従来の手法では，人手であらかじめ単語 N-gram，文字 N-gram や辞書などを設定した上で特徴量を抽出し，それらを素性として CRF を適用することで対象となる固有表現にタグを付与していくのが一般的であった．しかし，本節で紹介した手法では，深層学習を利用して特徴量を抽出し，それらを素性として CRF や他の深層学習を利用することで固有表現タグの付与を行っている．つまり，辞書などの外部知識を用いずに深層学習で獲得した素性を用いることで，固有表現抽出の最高性能を達成できるということになる．これは，タスクに依存した素性を人間が設計する必要が無くなったこと，文字や単語の埋め込みを学習することで特徴量の抽出を自動化できるようになったことを意味する．

しかし，これらの研究を含め，深層学習を用いた固有表現抽出に関する研究の大多数は CoNLL-2003[7]で使用された英語のデータセットを対象としており，最先端の固有表現抽出モデルが他の言語にも有効か明らかにされていなかった．しかし最近の研究では，英語以外を対象とした研究が見られるようになってきてお

り，Misawa ら[8]は日本語を対象に LSTM-CRF とその派生モデルの有効性を検証した上で，新たなモデルを提案している．Misawa らは，Ma らの提案モデルを日本語のテキストに対して適用することで検証を行った結果，以下の 2 つの障害を明らかにした．

- CNN による文字表現の抽出は，日本語には適していない．
- 単語の一部が固有表現を構成するときに、単語ベースのモデルは領域を抽出できない．

1 つ目の理由は単語の長さが英語よりも日本語の方が短くなる傾向があること，2 つ目の理由は日本語には明示的な単語区切りがなく，単語の境界があいまいになるため固有表現の境界が単語の境界と一致しない可能性があることが原因と述べている．(例：東京都内 → 東京/都内，東京都/内）これらを踏まえ，Misawa らは図 4 のモデルを提案している．

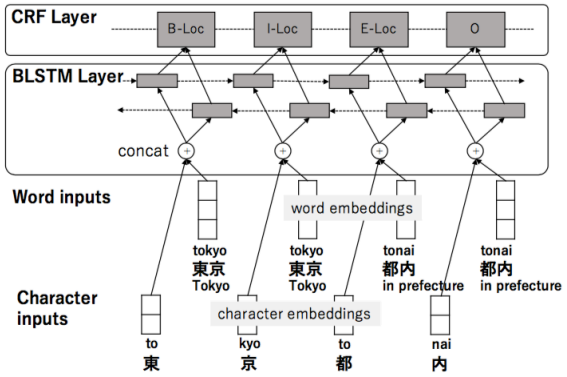

図 4：Misawa らの提案モデル
((Misawa, 2017[8])より抜粋)

このモデルは，文字埋め込みを入力として与えて文字の固有表現タグを推定するようになっている．また，文字を入力する際に単語埋め込みも入力して両者の情報を結合し，Bi-LSTM layer に入る仕組みになっている．このモデルは，日本語を対象とした固有表現抽出において最高性能を達成している．

本研究では，Lample らのモデルをベースとし，技術関連記事用に拡張したモデルを提案する．そのため，Lample らのモデルは 4 章で詳しく説明する．また，単語埋め込み，LSTM についても 4 章で詳しく説明を行う．また，本研究では，入力は単語埋め込みを用い，文字埋め込みは利用しない．

4. 経営判断支援システムの構築

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，技術関連記事から組織名などを抽出するモジュールから構成される．本節では，4.1 節でシステム概要，4.2 節で技術関連記事からの固有表現抽出，4.3 節で技術関連記事のカテゴリ分類について説明する．

4.1. システム概要

システム全体の処理手順は，我々の先行研究[9]に基づいて，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）を利用する．表 2 に Cls→NE 法を実現する上で用いる手法，図 5 にシステム動作の流れを示す．

表 2：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[10]	Bi-directional LSTM-CRF
	階層型 3 層ニューラルネットワーク	Bi-directional LSTM-CRF

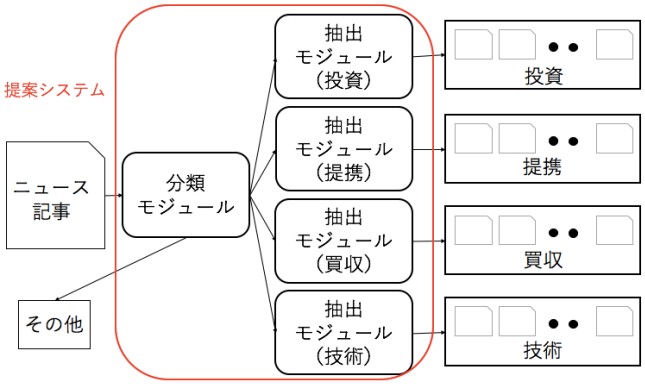

図 5：システム概要（Cls→NE 法）

4.2. 技術関連記事からの固有表現抽出

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モジュールを構築する．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 6，図 7，図 8，図 9 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
…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 6：「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。…

図 7:「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 8:「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 9:「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，深層学習に基づく固有表現抽出手法である Bi-directional LSTM-CRF（以下 Bi-LSTM-CRF）を用いる．本研究で取り扱う Bi-LSTM-CRF モデルを図 10 に示す．

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）

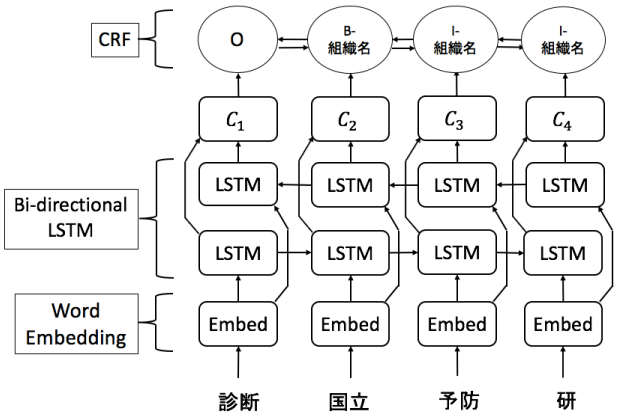

図 10：本研究で扱う Bi-LSTM-CRF モデル
（Lample et al.(2016)[5]の Figure 1 を参考に作成）

図 10 のモデルは，Lample らのモデルをベースとしている．このモデルは Word Embedding，Bi-directional LSTM(以下 Bi-LSTM)，CRF の 3 層からなる．本節では，このモデルについて詳しく説明する．なお，本研究では Misawa らの報告に基づき，CNN layer は使用しない．また，Misawa らは Word Embedding の他に Character Embedding を利用しているが，本研究では分類モジュールと固有表現抽出モジュールの入力単位を同一にするため，Character Embedding を使用しない．

Word Embedding

まず，入力は単語であり，Word Embedding にて単語埋め込みに置き換えられる．単語埋め込みとは，word2vec などの手法を用いて単語を 300 次元程度の実数ベクトルで表現する技術のことを指し，現在の自然言語処理において重要な技術になっている．この技術により，意味の近い単語をベクトルによって対応させることが可能な他，ベクトルの演算によって意味のある結果が得ることが可能になっている．(例 : king - man + women = queen)

Bi-directional LSTM

次に，Bi-LSTM に単語埋め込みが入力される．LSTM は時系列データを扱うことができ，入力ゲート$\boldsymbol{i}$，忘却ゲート$\boldsymbol{f}$，出力ゲート$\boldsymbol{o}$の 3 つのゲートを利用することで長期の記憶を行うニューラルネットワークである．また，長期の情報を保持するセル（cell）と呼ばれる隠れ状態ベクトル$\boldsymbol{c}$，次の層や出力層で利用される隠れ状態ベクトル$\boldsymbol{h}$を用いる．それぞれの値の計算式は以下の通りである．⊙は要素ごとの積を表している．

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i\boldsymbol{x}_t + \boldsymbol{W}_i\boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \quad (4.1)$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f\boldsymbol{x}_t + \boldsymbol{W}_f\boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \quad (4.2)$$

$$\boldsymbol{c}_t = \boldsymbol{i}_t \odot (\boldsymbol{W}_f\boldsymbol{x}_t + \boldsymbol{W}_f\boldsymbol{h}_{t-1}) + \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} \quad (4.3)$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o\boldsymbol{x}_t + \boldsymbol{W}_o\boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \quad (4.4)$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t) \qquad (4.5)$$

入力ゲート$\boldsymbol{i}$で，t 番目の入力データ$\boldsymbol{x}_t$を基に t-1 番目の隠れ状態ベクトルの値を調整し，長期の情報を集約したセルである$\boldsymbol{c}_t$を更新する．さらに，忘却ゲートで過去のセルの値を減少させる．最後に，更新された$\boldsymbol{c}_t$の値と出力ゲート$\boldsymbol{o}_t$で調整した値を利用することで，最終的な隠れ状態ベクトル$\boldsymbol{h}_t$を生成する．

図 10 のモデルでは，現在の時間ステップよりも前に入力された情報を記憶しておくことができるという利点を生かし，周辺にどのような単語があるかを LSTM によって情報を集約する．また，順方向に解析する LSTM だけではなく逆方向に解析する LSTM を組み合わせることで，現在の時間ステップの前後情報を利用することができる．これにより，入力データの前後情報から素性を自動で獲得することが可能になる．

CRF

Bi-LSTM で獲得した素性を用いて，入力の単語に対して固有表現タグの付与を行う．本研究では，BIO タギングにより解決を試みる． BIO タギングとは，固有表現を単語などの単位に区切った際に，先頭の構成要素（Begin），2 番目以降の構成要素（Inside），固有表現を構成する要素でないもの（Other）の 3 タイプを考慮し，固有表現を表す手法である．図 10 のモデルの動作例では，「国立予防研」が 1 つの組織名として考慮され，先頭要素の「国立」に「B-組織名」，2 番目以降の要素に「I-組織名」が付与される．

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，1 つのニュース記事は 1 つのカテゴリにだけ分類されるようにする．分類モジュールは 2 種類の 3 層ニューラルネットワークに基づく手法を用いてそれぞれ構築する．1 つは階層型 3 層ニューラルネットワーク，もう一方は Joulin らが提案している fastText を用いる．

階層型 3 層ニューラルネットワーク

本研究では，図 11 のような階層型 3 層ニューラルネットワークを利用する．一番下が入力層になっており，文書を単語列$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$で表現したものが入る．この単語列は，Bag-of-Words という形式に変換したものを用いる．Bag-of-Words は，ある文書における単語の出現回数を数え，単語の並びを考慮せずに文書に単語が含まれているかどうかを考慮するモデルのことを指す．出力層で softmax によって各文書が各カテゴリに所属する確率分布$\boldsymbol{p} = (p_1, p_2, p_3, p_4, p_5)$を得る．その後，$\boldsymbol{p}$の中で最も確率値が高いカテゴリをその文書のカテゴリとして割り当てる．本研究で取り扱うニューラルネットワークは，本研究で定義した 5 カテゴリのいずれかに分類するように学習する．

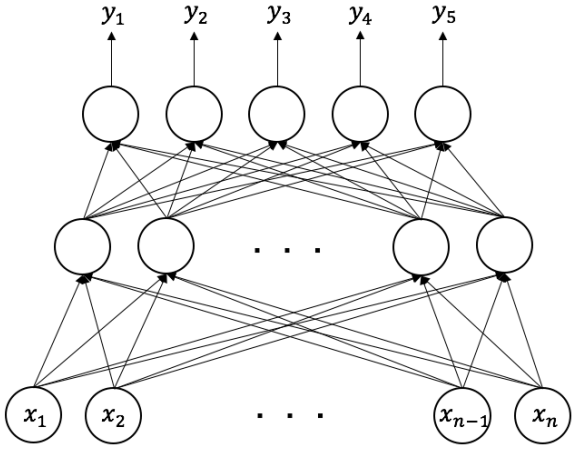

図 11：階層型 3 層ニューラルネットワークに基づく分類モデル

fastText

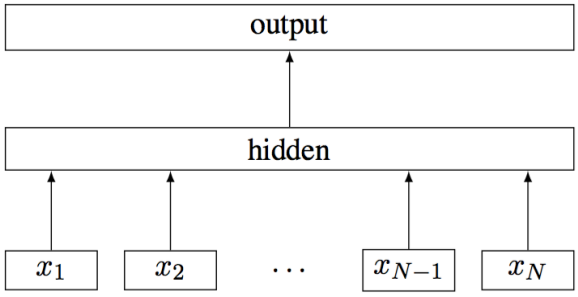

図 12：fastText（Joulin et al.(2016)[10]より抜粋）

Joulin らが提案している fastText のモデルを図 12 に示す．fastText は単語埋め込み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．Bojanowski ら[11]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」を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．このモデルの入力$\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$は，文書に含まれる単語の one-hot ベクトルとなる．出力層で softmax によって各文書が各カテゴリに所属する確率分布$\boldsymbol{p}$を得て，最も確率値が高いカテゴリを入力文書に割り当てる．階層型 3 層ニューラルネットワークとの違いは以下の通りである．

- fastText
  単語埋め込みにより，単語の類似度が考慮され，意味関係を捉えることができる．

- 階層型 3 層ニューラルネットワーク
  Bag-of-Words により,単語が出現しているかと出現頻度を捉えることができる.

5. 実験

5.1. 技術関連記事からの固有表現抽出

毎日新聞，日本経済新聞，読売新聞から人手で技術関連記事を収集し，表 1 のカテゴリに分類を行った．次に，表 3 で示した固有表現タグを人手で付与した．表 4 に実験データとして用いる技術関連記事の件数，付与した固有表現抽出タグの件数を示す.

表 4：5.1 節の実験で使用する記事データ

	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	141	-	96	95
提携	104	401	-	-	-
買収	203	278	264	-	51
技術	155	334	-	-	-

実験データに対して 4.2 節で説明した Bi-LSTM-CRF を適用し，評価を行った．本実験では学習済みモデルとして，毎日新聞，日本経済新聞，読売新聞合計 65 年分の新聞記事データを word2vec によって学習させたものを用いる．また，本実験で設定したハイパーパラメータは以下の通りである.

・バッチサイズ: 5
・dropout: 0.5
・単語埋め込みの次元数: 300
・LSTM のユニット数: 300
・epoch: 100

比較手法には機械学習手法の一つである CRF を用いる．CRF においては，技術関連記事に適した抽出器にするため，人手により素性の選定を行った．素性は以下の通りである.

- ターゲットの形態素から,前後 3 形態素のユニグラム，バイグラム，トライグラム
- ターゲットの形態素から，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

深層学習に基づいた従来研究では，訓練，開発，テストデータの 3 つを用意し，訓練データを用いてモデルを作成した後に開発データを使ってハイパーパラメータをチューニングし，最終的にテストデータを使って評価するのが一般的である．しかし，本研究で取り扱うデータのサイズが従来研究におけるデータサイズより小規模であるため，矢野ら[12]の評価方法に基づき，CoNLL-2000 共有タスクと同じである n 分割交差検定により,評価を行う.本研究では予備実験の結果，n=2 とした．本研究では従来研究の評価手法に基づいて，評価尺度は精度，再現率，F 値を用いた.

5.2. 技術関連記事のカテゴリ分類

実験データには，表 4 で示した記事データを固有表現タグがついていない状態にし，「その他」の記事を 128 件追加したデータ集合を用いた．実験データに対して 4.3 節で説明した階層型 3 層ニューラルネットワーク，fastText を適用し，評価を行った．また，予備実験の結果，階層型 3 層ニューラルネットワークの中間層の次元数を 1000，epoch 数を 50 とし，fastText の epoch 数は 100 とした．本実験における比較手法として，Kim の CNN による文書分類手法を利用した．5 分割交差検定により評価を行い，評価尺度は精度，再現率を用いた.

6. 実験結果と考察

6.1. 固有表現抽出

5.1 節で説明した実験の結果を表 5 に示す．平均値に関して，有意水準 5%で t 検定を行ったところ，p=0.0131 となり，有意差を確認できた．本研究ではニュース記事からもれなく組織名などの技術関連の情報を抽出することが望ましいと考え,再現率を重視する.表 5 より，精度を見ると CRF が良く，再現率を見ると Bi-LSTM-CRF が良い性能を示した．また，CRF，Bi-LSTM-CRF 共に平均して 0.7 ポイント程度の F 値が得られているが，Bi-LSTM-CRF が 0.056 ポイント上回った．これより，再現率，F 値の観点から，技術関連記事においても CRF よりも Bi-LSTM-CRF の方が良いことが示された.

タグ別に結果を見ると，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が，投資対象の再現率が 0.365 ポイント，買収先が 0.489 ポイントと低い結果となった．買収においては買収先と買収元の区別を行っているが，両者とも組織名を扱うため，これらの区別が Bi-LSTM-CRF のような固有表現抽出手法では困難であることが原因と考えられる.

ここからは，投資対象タグにおいて再現率が 0.365 ポイントという結果になったことについてエラー例を図 13，図 14 に示し考察を行う.

正解：**<投資対象>**IT システムのための研究・開発**</投資対象>**
実験結果：IT システムのための研究・開発

図 13：Bi-LSTM-CRF におけるエラー例（投資）

投資対象では,「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,図 13 のように名詞以外の品詞が含まれる固有表現は正解と同じように抽出できていなかった例が見られた．投資対象の固有表現に対して全くタグが付与されない例の他には，図 14 のように正解の一部にしかタグ付与されない例も見られた．

正解：**<投資対象>**第六次空港整備五カ年計画**</投資対象>**の総投資額・・・

実験結果：**<投資対象>**第六次空港整備**</投資対象>**五カ年計画の総投資額・・・

図 14：Bi-LSTM-CRF におけるエラー例 2（投資）

図 14 の例では，正解では「第六次空港整備五カ年計画」に投資対象タグが付与されていることに対し，実験結果では「第六次空港整備」にだけ投資対象タグが付与されている．本実験ではタグが完全一致している場合のみ正解としているため，この例は不正解となるが，本来抽出すべき情報をすべて抽出できていなくても,「第六次空港整備」のように最低限投資対象として意味のわかる固有表現を抽出できていることがわかる．投資対象ではこのような例が 9 件見られた．

表 5：固有表現抽出モジュールの実験結果

		CRF			Bi-LSTM-CRF		
	タグ	精度	再現率	F 値	精度	再現率	F 値
投資	金額	**0.780**	0.821	0.800	0.764	**0.853**	**0.806**
	組織名 投資元	**0.884**	0.702	0.783	0.876	**0.801**	**0.837**
	投資 対象	**0.609**	0.146	0.236	0.393	**0.365**	**0.378**
提携	組織名	**0.890**	0.771	0.826	0.871	**0.878**	**0.875**
買収	金額	0.648	0.686	0.666	**0.722**	**0.765**	**0.743**
	組織名 買収元	0.728	0.597	0.656	**0.748**	**0.748**	**0.748**
	組織名 買収先	0.565	0.310	0.400	**0.662**	**0.489**	**0.562**
技術	組織名	**0.845**	0.752	**0.796**	0.813	**0.755**	0.783
平均	-	**0.789**	0.621	0.695	0.774	**0.730**	**0.751**

6.2. カテゴリ分類

5.2 節で説明した実験の結果を表 6 に示す．表 6 から分かる通り，比較手法である CNN が 0.365 ポイントという結果となったが，fastText で 0.883 ポイント，階層型 3 層ニューラルネットワークで 0.897 ポイントという結果が得られた．fastText，階層型 3 層ニューラルネットワークの実験結果について，有意水準 5%で t 検定を行ったところ p=0.2407 となり，有意差を確認できなかった．そこで本節では，fastText と階層型 3 層ニューラルネットワークの結果を分析し，考察を行う．表 7 に本実験結果をカテゴリ別に示す．

表 6：カテゴリ分類モジュールの実験結果

	手法	精度	再現率
Cls→NE 法	fastText	0.883	0.883
	階層型 3 層 NN	0.897	0.897
	CNN	0.365	0.365

表 7：カテゴリ別の実験結果

	fastText			階層型 3 層 NN		
	精度	再現率	F 値	精度	再現率	F 値
投資	0.862	**0.909**	0.885	**0.949**	0.845	**0.894**
提携	**0.952**	0.914	**0.933**	0.826	**0.962**	0.889
買収	0.910	0.941	0.925	**0.951**	**0.956**	**0.953**
技術	**0.902**	0.948	**0.925**	0.876	**0.961**	0.917
その他	0.775	0.665	0.716	**0.860**	**0.719**	**0.783**
全体	0.883	0.883	0.883	**0.897**	**0.897**	**0.897**

カテゴリ別の内訳を F 値で比較すると，提携，技術で fastText が階層型 3 層ニューラルネットワークを上回り，投資，買収，その他では階層型 3 層ニューラルネットワークが fastText を上回った．精度，再現率で比較すると，その他と買収において階層型 3 層ニューラルネットワークが上回った．しかし，両手法において，その他への分類性能が低い傾向が見られ，fastText では再現率が 0.7 ポイントを下回る結果となった．以下に fastText における，その他の記事の誤分類例を図 15 に示す．図 15 の記事は「国際協力銀行」と「インドステイト銀行」の提携に関する記事のため，正解カテゴリは提携となる．しかし，この記事では「銀行」,「投資」,「融資」のように投資に頻出するキーワードが多く存在しており,単語の類似度を考慮する fastText で投資に誤分類されたと考えられる．階層型 3 層ニューラルネットワークでは，単語の表層情報を考慮するため，分類する上で重要になる単語は fastText よりも少ないと考えられる．つまり，図 15 のような例では，fastText の単語埋め込みによる情報がかえってノイズになり，誤分類されたと考えられる．

これより，fastText は単語間の意味関係を捉えるため，カテゴリの特有のキーワードと類似している単語が多く出現していると，そのカテゴリになりやすい傾向があると考えられる．技術関連記事では、カテゴリ名を表す単語（投資，提携，買収，技術）がそもそも分類する上で最大のキーワードとなり，この単語があるかないかだけでも分類がある程度可能と言える．これより，階層型 3 層ニューラルネットワークで得られる単語の表層情報，出現頻度が分類する上で fastText より有効であったと考えられる．特に，その他の記事に関してはこの傾向が特に見られ，記事をその他か技術関連記事の 4 カテゴリに分類する上で階層型 3 層ニューラルネットワークが有効と言える．

（2006/9/20 読売新聞）

国際協力銀：印企業に７０００万ドル融資枠、現地銀と提携　輸出活性化狙い

国際協力銀行は１９日、インド最大の国営商業銀行、インドステイト銀行と提携し、０７年から３年間で計７０００万ドル（約８２億円）の融資枠を設定する方針を明らかにした。日本から機械設備などを輸入する現地企業を対象に融資する。・・・

国際協力銀が日本企業を対象に行った投資期待先調査でインドは中国に次ぎ２位。年８％前後の高い経済成長を続けていることもあり、日本企業の投資意欲は年々強まっている。

大型の融資枠の設定で、現在は対中輸出額の２０分の１以下にとどまっている日本からの対インド輸出を一層活性化させる狙いだ。

両行の関係者が２０日、シンガポールで調印する。インドステイト銀行の支店を通じて隣国のスリランカ企業にも門戸を広げる計画で、活用する企業が多ければ、融資期間の延長も検討するという。

図 15：誤分類例
（正解：提携，fastText：投資，
階層型 3 層ニューラルネットワーク：提携）

階層型 3 層ニューラルネットワークの問題点として，入力の単語列を Bag-of-Words 形式に変える際に利用する辞書の単語数が増えると，それに応じて入力単語列も高次元になってしまうことが挙げられる．fastText では，辞書の単語数が増えても単語埋め込みを利用することから低次元ベクトルのままで済むため，実際のシステムで利用する場合は fastText が優れていると考えられる．そのため，将来的に技術関連記事のデータセットの規模が大きくなった際には，fastText の利用を検討する必要がある．

7. おわりに

本研究では，深層学習に基づく手法を用いて，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を行い表にまとめるシステムを構築した．固有表現抽出の実験では，F 値において Bi-LSTM-CRF が F 値で 0.751 ポイントを得ることができ，CRF を上回ったことで有効性を確認することができた．カテゴリ分類実験では，階層型 3 層ニューラルネットワークが精度・再現率ともに 0.897 ポイントを得ることができ，性能，コスト面から考察を行った結果，有効性が確認できた．最後に，本研究で構築した経営判断支援システムにより，図 1，図 2 のような表形式で技術関連記事の重要な情報の表示を行った．このシステムで技術分野の動向を概観できることにより，経営判断の一助になると考えられる．

参　考　文　献

[1] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. arXiv:1408.5882.

[2] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781v3 [cs.CL].

[3] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. J., and Hovy, E. H. (2016). Hierarchical Attention Networks for Document Classification. In HLT-NAACL, 1480-1489.

[4] Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML Vol. 1, 282-289.

[5] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural Architectures for Named Entity Recognition. arXiv:1603.0136v3 [cs.CL].

[6] Ma, X., and Hovy, E. (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. arXiv:1603.01354v5 [cs.LG].

[7] Sang, E. F. T. K., and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Vol. 4, 142-147. Association for Computational Linguistics.

[8] Misawa. S., Taniguchi, M., Miura, Y., and Ohkuma, T. (2017). Character-based Bidirectional LSTM-CRF with Words and Characters for Japanese Named Entity Recognition. In Proceedings of the First Workshop on Subword and Character Level Models in NLP, 97-102.

[9] 平前 歩, 乗重 雅誉, 難波 英嗣, 竹澤 寿幸. (2017). 技術関連記事の自動要約．第 9 回データ工学と情報マネジメントに関するフォーラム (DEIM 2017).

[10] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. arXiv:1607.01759v3 [cs.CL].

[11] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching Word Vectors with Subword Information. arXiv:1607.04606v1 [cs.CL].

[12] 矢野 憲, 伊藤 薫, 若宮 翔子, 荒牧 英治. (2017). 深層学習による医療テキストからの固有表現抽出器の開発とその性能評価．第 31 回人工知能学会全国大会(JSAI2017)，2J2-OS-16a-4.