技術関連記事の自動要約

Automatic Summarization of Technical News Articles

平前　歩
Ayumu Hiramae

広島市立大学大学院　情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

概要 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．企業の経営者が経営判断をする際，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が必要となる．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．

1　はじめに

日々新技術や製品の開発を行っている企業の経営者にとって，他の企業がどのような技術を発明したか，大学と提携して技術開発を行っているか，といった技術関連の情報を入手することは経営判断をする上で重要である．それらの情報を入手するためには，新聞記事，オンラインニュース記事などが役立つ．しかし，それらのニュース記事は技術関連の情報以外にも多数の情報が存在し，目的の情報を把握するのに時間がかかってしまう．そこで本研究では，ニュース記事の種類を自動で判別する分類器を構築し，その種類に応じて異なる要約手法を適用することで，より適切な要約を作成する．

2　関連研究

McKeown ら[1]のグループは，CNN 等のニュースサイトから新聞記事を収集し，その新聞記事をイベントごとにクラスタリングを行い，さらにイベントクラスタごとに要約を生成する Newsblaster というシステムを構築している．Newsblaster では，イベントクラスタのタイプを考慮し，タイプによって異なる要約方法を用いている．Mckeown らは複数の新聞記事からなるクラスタのタイプによって要約手法を変えているが，本研究では 1 つのニュース記事の種類を判別し，その種類によって考慮した要約を作成する．

3　ニュース記事の自動要約

3.1　システム概要

本研究で構築するシステムについて，図 1 を用いて説明する．システムに入力したニュース記事は，カテゴリ分類器にて 8 つのカテゴリのいずれかに分類される．その後，カテゴリに応じた要約システムを用いて要約を作成し，出力する．図 1 は，入力のニュース記事がカテゴリ分類器によって「投資」に分類され，「投資」用の要約システムによって要約が出力される場合の例である．本研究では技術関連のニュース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．

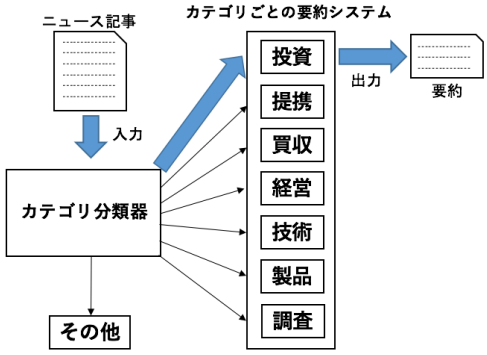

図 1：システムの構成と動作例

3.2　カテゴリ分類と固有表現抽出

本研究では，ニュース記事を以下のカテゴリに分類する．

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
経営：企業の経営体制・方針
技術：新しく開発した技術
製品：新しい技術による製品開発，改良
調査：新しい技術の研究，調査の着手
その他：上記以外，一般論（論説，ブログ等）

まず，人手でニュース記事を以上の 8 カテゴリに分類し，その他以外の各カテゴリにおいてニュース記事に含まれる重要な単語に対してタグ付けを行った．「買収」，「提携」それぞれの例を図 2，図 3 に示す．

<org>ナブコ</org>は十三日、米国の中堅自動ドアメーカーの<tr>ランソン・インダストリーズ社</tr>(本社・ウィスコンシン州)と<tr>グループ二社</tr>を、<money>約十億円</money>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。同社の全株式(二千七百四十五株)を取得したもので、経営陣はそのまま残し、社員一人を副社長として派遣する。

図 2：「買収」記事

図 2 の<org></org>は買収元の企業名，<tr></tr>は買収先の企業名，<money></money>は金額を表している．

ネット銀行業務への参入を予定していた<org>楽天</org>の提携先が、<org>東京都民銀行</org>に固まったことが１８日、明らかになった。今年夏をめどに、ネット上に都民銀の仮想支店を開いて実績を積んだ後、来年にも共同でネット専門銀行を設立する方針という。楽天はまた、証券取引業で既に提携している新生銀行と、住宅ローン専門会社を共同で設立する方針も固めた。既に損害保険業務への参入も決めており、総合ネット金融企業としての陣容が一層整うことになる。

図 3: 「提携」記事

図 3 の<org></org>は提携する 2 つの企業名を表している．これらのタグがついた単語を含む文を要約に加えることで，より良い要約が作成できると考えられる．本研究では，ニュース記事における重要な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．また，カテゴリによって抽出すべき固有表現が違ってくると考えられるため，カテゴリそれぞれに適した固有表現抽出器を作成する必要がある．

3.3 要約手法

この節では，カテゴリに応じた要約システムにおいて利用する代表的なテキスト自動要約手法について述べる．

LexRank

Erkan ら[2]は文をノード，文間の関係をエッジとした類似度グラフを作成し，そのグラフを基に文の重要度を計算する手法を提案している．図 4 は類似度グラフの例である．LexRank は，単に次数の多いノードを評価するだけでなく，次数の多いノードと隣接しているノードの重要度についても考慮している．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．

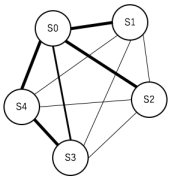

図 4：類似度グラフの例

ILP(Integer Linear Programming)

Woodsend ら[3]は，テキスト自動要約を最適化問題の一つである整数線形計画問題に定式化している．ニュース記事の中の重要な文を選択して要約を含めるのではなく，フレーズ単位で選択して要約を作成する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する．

Deep Learning

Nallapati ら[4]は，機械翻訳，音声認識の分野で成功を収めている Deep Learning をテキスト自動要約に適用している．彼らは，機械翻訳に使われることが多い RNN（Recurrent Neural Network）を用いたシーケンス変換モデル（sequence to sequence model）をベースにニュース記事要約モデルを構築している．

4 デモシステム動作例

本研究で構築したデモシステムについて説明する．例として，図 5 のニュース記事を入力とした場合の出力例を図 6 に示す．本システムでは ILP による要約モデルを用いている．入力のニュース記事の要約だけでなく，そのニュース記事における重要語を可視化する．例のような「買収」に関する記事の場合は，金額，買収先，企業名（買収元）のそれぞれを色で分けて表示する．

米軍事関連大手のマーチン・マリエッタとグラマンは七日、合併することで合意したと発表した。マーチン側が、グラマン株を総額十九億ドルで買収する。東西冷戦の終結に伴って軍事産業に対する需要が大幅に落ち込んだのに対応するためで、米政府の認可を得て合併が実現すれば、年間売上高百三十億ドルにのぼる米最大規模の軍事企業が誕生することになる。今回の合併についてグラマンのレンソー・キャポラリ会長は「現在のビジネスの状況では大きな戦略的な動きをしない限り、会社の繁栄はないと一年前に決意し、いろいろ手段を検討してきたが、マーチン社との合併が最良と判断した」と説明している。米軍事産業は、冷戦の終結に伴い積極的にリストラ（事業の再構築）に取り組んでいるが、状況は依然として厳しく、今後大規模な再編が進む可能性もある。

図 5：システム入力例

情報抽出結果

米軍事関連大手のマーチン・マリエッタとグラマンは七日、合併することで合意したと発表した。マーチン側が、グラマン株を総額十九億ドルで買収する。東西冷戦の終結に伴って軍事産業に対する需要が大幅に落ち込んだのに対応するためで、米政府の認可を得て合併が実現すれば、年間売上高百三十億ドルにのぼる米最大規模の軍事企業が誕生することになる。今回の合併についてグラマンのレンソー・キャポラリ会長は「現在のビジネスの状況では大きな戦略的な動きをしない限り、会社の繁栄はないと一年前に決意し、いろいろ手段を検討してきたが、マーチン社との合併が最良と判断した」と説明している。米軍事産業は、冷戦の終結に伴い積極的にリストラに取り組んでいるが、状況は依然として厳しく、今後大規模な再編が進む可能性もある。

金額 買収先 企業名(買収元)

要約

米軍事関連大手のマーチン・マリエッタとグラマンは七日、合併することで合意したと発表した。
マーチン側が、グラマン株を総額十九億ドルで買収する。
米軍事産業は、冷戦の終結に伴い積極的にリストラ（事業の再構築）に取り組んでいるが、状況は依然として厳しく、今後大規模な再編が進む可能性もある。

図 6：システム出力例

5 おわりに

本研究では，ニュース記事を機械学習によってカテゴリ分類し，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．今後の課題として，カテゴリごとにより精度の高い固有表現抽出器を作成する必要がある．また，Deep Learning による要約システムの構築を行い，評価を行う予定である．その上でカテゴリごとに適切な要約手法を検討する．

6 参考文献

[1] McKeown, K. R., et al. (2002). Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster. Proceedings of the Second International Conference on Human Language Technology Research.

[2] Erkan, G., et al. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research, 22, 457-479.

[3] Woodsend, K., et al. (2010). Automatic Generation of Story Highlights. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 565-574.

[4] Nallapati, R., et al. (2016). Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. arXiv:1602.06023v2 [cs.CL].