決算短信からの事業セグメント情報抽出

伊藤 友貴 [1,2] *　　小林 暁雄 [2]　　関根 聡 [2]

[1] 東京大学大学院工学系研究科　　[2] 理研 AIP

m2015titoh@socsim.org, {akio.kobayashi, satoshi.sekine}@riken.jp

1 はじめに

情報通信技術の発達に伴い，金融テキストマイニングの技術に対する投資家からの関心が高まってきている．投資家が投資活動を行うにあたり，上場企業の業績情報の収集は必要不可欠である．その中でも上場企業が四半期毎に公開される「決算短信」は投資判断を行う上で有用な情報源の一つである．決算短信からの情報抽出に関する研究はいくつか行われており，関連研究として企業の決算短信から業績要因文を自動抽出する研究 [1] や決算短信からの業績文要約生成を行う研究 [2] が挙げられる．本研究では特に事業セグメントの情報の決算短信からの抽出，すなわち「事業セグメント名」及び「事業セグメントの内容」の抽出に焦点を当てる．事業セグメントの情報は決算短信の解析をする上で重要な情報である．各企業は多くの場合，図 1 に挙げた例のように事業セグメント（企業の構成単位）毎に決算情報を記述する．さらに,「事業セグメント名」及び「事業セグメントの内容」の情報と「事業セグメント別の売上情報」を組み合わせることで各企業の経営状態の可視化などができる可能性がある．このように,「事業セグメント名」及び「事業セグメントの内容」の情報は投資判断の上で重要である．しかし,「事業セグメント名」及び「事業セグメント内容」の決算短信からの自動抽出手法については確立されていない．また，現状，決算短信特有の言語的な特徴に関する知見はあまり共有されていない．そこで，本研究では決算短信特有の言語的な特徴を考慮した「事業セグメント情報抽出手法」を提案し（2 節），その有用性について実データを用いて評価する（3 節）．

> 当社の報告セグメントは、当社の構成単位のうち分離された財務諸表が入手可能であり、取締役会が、経営資源の配分の決定及び業績を評価するために、定期的に検討を行う対象となっているものであります。当社は、製品・サービス別に「水産商事」、「冷凍食品」、「常温食品」、「物流サービス」、「鰹・鮪」、「その他」を報告セグメントとしております。「水産商事」は、冷凍魚介類等水産物の買付、販売を行っております。「冷凍食品」は、冷凍食品の製造、買付、販売を行っております。…

図 1: 決算短信の文面例 (「極洋」の平成 23 年第一四半期に公開された決算短信より抜粋).

*本研究は伊藤が理研 AIP にてインターンシップをしていたときの仕事である．

2 事業セグメント情報抽出

2.1 問題設定

本研究では訓練用データとして決算短信，それに記載される「事業セグメント名」及び「事業セグメント説明」が与えられており，それらの情報を用いて検証用データとして提供される決算短信から「事業セグメント名」,「事業セグメント説明」を抽出することを目指す．また，訓練用データとして提供される決算短信の企業と検証用データとして提供される決算短信の企業には被りがないという状況下での処理を想定する．

2.2 事業セグメント情報抽出手順概要

本研究では以下の Step 1 から Step 5 からなる決算短信から「事業セグメント名」及び「事業セグメント説明」を抽出する手法を提案する．

Step 1 決算短信から「セグメント情報」を記載している段落を抽出.

Step 2 事業セグメント名抽出 (2.3 項).

Step 3 事業セグメント名を含む文を抽出.

Step 4 Step 3 で抽出した文から不要な文を取り除き，各事業セグメントの説明を記載している文を抽出 (2.4 項).

Step 5 Step 4 で抽出した文から各事業セグメントの内容部分 (= 事業セグメント説明) を抽出 (2.5 項).

具体例として 図 1 に挙げたセグメント情報記載段落について考える．まず，Step 2 で，事業セグメント名「水産商事」、「冷凍食品」、「常温食品」、「物流サービス」、「鰹・鮪」、「その他」を抽出する．

次に Step 3 で,「水産商事」セグメントについて考える場合,「水産商事」を含む以下の二文を抽出する．

- 当社は、製品・サービス別に「水産商事」、「冷凍食品」、「常温食品」、「物流サービス」、「鰹・鮪」、「その他」を報告セグメントとしております。
- 「水産商事」は、冷凍魚介類等水産物の買付、販売を行っております。

次に Step 4 で上記の二つの文のうち事業セグメントの説明を含まない文を除き，以下の

- 「水産商事」は、冷凍魚介類等水産物の買付、販売を行っております。

という「水産商事」の内容を含む文のみを抽出する．

最後に Step 5 によって「水産商事」の説明が「冷凍魚介類等水産物の買付、販売を行っております。」であると判断する．今回の例では一文の中に単一の事業セグメントの情報しか記載されていないので Step 5 は容易であるが，実際には図 2 の例のように一文の中に複数の事業セグメント (「建設事業」及び「不動産事業等」) の情報が記載されている場合もある．この場合に「建設事業」の内容が「建設工事全般に関する事業」であり,「不動産事業等」の内容が「不動産の売買及び賃貸等に関する不動産事業等」であると判断することが Step 5 において求められる．

上記の中で難しいのは Step 2, Step 4, Step 5 である．そこで，本研究ではこれらの手順を行う手法を決算短信の持つ言語的性質を考慮して提案する．以下，Step 2，Step 4，Step 5 における各手法について具体的に説明する．

2.3 事業セグメント名抽出 (Step 2)

本研究にて提案する事業セグメント名抽出手法は以下の Step 2.1 〜 Step 2.3 から成る．

Step 2.1 事業セグメント情報段落内の各単語に以下の手順 (Bidirectional Bag of words (*BBOW*)) で単語ベクトルを与える．

- 各単語の直前 10 単語に出現する単語の頻度，及び各単語の直後 10 単語 10 単語に出現する単語の頻度によってベクトル v_f, v_b を生成する．
- $[v_f, v_b]$ を単語ベクトルとして与える．

Step 2.2 訓練データ内の単語について事業セグメント名であれば正，そうでなければ負のラベルを与え，訓練データ内の単語を用いて予測モデル (Logistic 回帰モデル) を構築しする．

Step 2.3 検証データ内の単語について学習済みの予測モデルを用いて正か負かの判断を行い，各単語が事業セグメント名であるかどうかを判断する．

ここで，Step 2.1 における特徴量生成法 *BBOW* は以下に挙げる本タスク特有の性質を考慮した，本手法特有の特徴量エンジニアリング手法である．

- ラベルが負の単語数がラベルが正の単語数の 50 倍以上であるというインバランスなデータを扱っており，高い Precision で抽出する必要がある．
- 各企業が似た言い回しでセグメント情報を記述する．
- 各単語について単語の前に出現する単語の頻度分布と単語の後に出現する単語の頻度分布は異なる．

2.4 事業セグメント説明文抽出 (Step 4)

本研究にて提案する事業セグメント説明文抽出手法は次の通りである．訓練データセット内の事業セグメントを含む文についてセグメント説明を含む文を正例，含まない文を負例としてラベルを与え，訓練データセット内の文を用いて予測モデル (Logistic 回帰モデル) を学習させる．その後，検証データ内の事業セグメントを含む文について学習済みの予測モデルを用いて正か負かの判断を行い，セグメント説明を含む文かどうかを判断し抽出する．このとき，文の素性には bag of words を用いる．

2.5 事業セグメント内容抽出 (Step 5)

2.5.1 事業セグメント情報記載文の性質

各事業セグメントに関する説明を記載している文は大きく以下の「単一型」,「Forward 型」,「Backward

型」の 3 タイプに分類できる．今回検証した 320 の決算短信中において，セグメント説明を含む文は 580 文見られた．このうち 450 が単一型，90 が Forward 型，27 が Backward 型であった．

単一型: 文中に単一の事業セグメントの情報のみ含むもの．

Forward 型: 一文中に複数の事業セグメントの説明があり,「セグメント名」,「セグメント説明」の順でセグメント情報が記載されるもの (図 2).

Backward 型: 一文中に複数の事業セグメントの説明があり,「セグメント説明」,「セグメント名」の順でセグメント情報が記載されるもの (図 3).

> 「建設事業」は建設工事全般に関する事業を、「不動産事業等」は不動産の売買及び賃貸等に関する不動産事業等を行っております．

図 2: Forward 型の文の例（名工建設が 2011 年 3 月に公開した決算短信より抜粋).

> 当社グループは、調剤薬局、ジェネリック医薬品の販売、人材紹介業及びコンサルティング業等により構成される「医薬事業」、都市型、コスメ型、郊外型のドラッグストアの経営等により構成される「物販事業」を軸とし …

図 3: Backward 型の文の例 (アインホールディングス公開の決算短信より抜粋).

2.5.2 事業セグメント内容抽出手法

以上の決算短信特有の性質を踏まえて事業セグメント内容抽出手法を以下のように提案する．

Step 5.1 事業セグメント名を用いて Step 4 で抽出された文について「単一型」のものを抽出し，それを事業セグメント説明とする

Step 5.2 訓練データ中の文を用いて「Forward 型」と「Backward 型」を分類する予測モデル (Logistic 回帰モデル) を構築し,「単一型」の文以外の文について 予測モデルを用いて「Forward 型」か「Backward 型」に分類する．このとき，文の素性には bag of words を用いる.「Forward 型」と分類された文については各セグメント名の後ろに記載されている文字列を抽出することで，また,「Backward 型」と分類された文については各セグメント名の前に記載されている文字列を抽出することで各セグメントの説明部分を抽出する．

3 セグメント情報抽出手法検証

本節では 2.3 項，2.4 項 , 2.5 項にて提案された各手法の妥当性，及び第 2 節にて提案した事業セグメント情報抽出手法の妥当性を実データを用いて検証する．

3.1 データセット・前処理

「セグメント情報」を決算短信に記載している 320 企業の決算短信を対象に実験を行った．各企業から 1 つずつ決算短信を取り出し（計 320 決算短信)，本実験に使用した．これらの決算短信 からセグメント情報を記載している段落内のテキストを抽出し，それらを Mecab[4] を用いて分かち書きしたものを本節における実験に使用した．

3.2 セグメント名抽出検証

本実験では 2.3 項におけるセグメント名抽出手法が妥当であるかどうかを検証した．

3.2.1 データセット・実験設定

320 決算短信中 250 企業を訓練データ，残り 70 企業を検証データとして使用し，2.3 節で紹介した手法によってセグメント名の抽出ができるかどうかを検証した．訓練データ，検証データそれぞれに含まれる事業セグメント名の数 （正例の数）は 713, 191 であり，事業セグメントではない単語の数（負例の数）は 40687，10466 であった．

また，Step 2.1 において用いた特徴量生成手法 *BBOW* の有用性を検証するため，Step 2.1 における単語ベクトルの与え方を以下の 3 つの特徴量生成手法にした場合の結果と比較した．

BOW: $v_f + v_b$ を単語ベクトルとして与える．

Skipgram: skipgram [3] によって単語に分散表現を与え，単語ベクトルとして与える．skipgram の学習コーパスには 320 企業の決算短信全体に含まれる文全てを用いた．

Skipgram 平均 (Bidirectional): 各単語に以下の手順で単語ベクトルを与える．まず，各単語の直前 10 単語に出てくる単語の分散表現の平均値 w_f ，及び各単語の直後 10 単語に出てくる単語の分散表現の平均値 w_b を計算し，$[w_f, w_b]$ を単語ベクトルとして与える．

3.2.2 実験結果・考察

本実験の結果（正例抽出の Precision, Recall, F-measure の値）は表1の通りである．これより，*BBOW* によりある程度うまくセグメント名の抽出ができること，及び決算短信特有の言語的性質を活かした手法 *BBOW* が他に比べ有用であることが確認できた．

表 1: 事業セグメント名抽出検証実験結果

手法	Precision	Recall	F-measure
BBOW	**1.00**	0.85	**0.92**
BOW	0.33	0.76	0.46
Skipgram	0.26	**0.98**	0.41
Skipgram 平均 (Bidirectional)	0.41	0.95	0.57

3.3 事業セグメント説明文抽出検証

2.4 項で紹介した手法の妥当性を検証した．検証にはデータセット内のセグメント名を含む文のうち，セグメントの説明を含む正例の文 580 件と含まない負例の文 870 件を用いた．検証の結果，Macro F 値（5 交差検定平均）0.870 で分類できることが確認でき，本手法の妥当性を確認できた．

3.4 事業セグメント内容抽出検証

2.5 項 の Step 5.2 で紹介した手法により Forward 型と Backward 型の分類がどの可能なのかについて検証した．検証にはデータセット内の Forward 型の文 90 文，Backward 型の文 27 文を用いた．検証の結果，Macro F 値 （5 交差検定平均） 0.892 で分類できることが確認でき，本手法の妥当性を確認できた．

3.5 事業セグメント情報抽出検証

最後に，Step 1 から Step 5 までを行った場合にどの程度事業セグメント名，及び事業セグメント説明を抽出できるかを検証した．訓練用データとして 288 企業の決算短信，検証用として 32 企業の決算短信を用いた．検証用データに含まれる事業セグメント名は 79 件，そのうち事業セグメント説明の記載がある事業セグメントは 65 件であった．事業セグメント名抽出の結果は Precision が $68/68 = 1.0$，Recall が $68/79 = 0.83$ という結果であった．また，事業セグメント説明抽出の結果は Precision が $35/48 = 0.73$, Recall が $35/65 = 0.54$ という結果であった．

4 結論

本研究において決算短信から事業セグメント名及びその説明を決算短信特有の言語的な特徴を用いて抽出する手法を提案し，実データを用いて評価を行った．評価の結果，本研究で提案した BBOW を用いた事業セグメント名抽出手法により高い精度で事業セグメント名を抽出可能であることが検証できた．その一方で本研究で提案した事業セグメント説明抽出手法は Precision が 0.73, Recall が 0.54 とやや課題の残る結果となった．今後の課題としては事業セグメント説明抽出手法の改善法の提案，BBOW が情報抽出タスクにおいて有用な条件の調査，及び「事業セグメント説明」,「事業セグメント名」を用いた企業が力を入れている事業の可視化手法の構築が挙げられる．

参考文献

[1] 酒井浩之，西沢裕子，松並祥吾，坂地泰紀, ”企業の決算短信からの業績要因の抽出”, 人工知能学会論文誌, vol. 30, no. 1, pp. 172–182, 2015.

[2] Isonuma, M., Fujino, T., Mori, J., Matsuo, Y., Sakata, I.: Extractive Summarization Using Multi-Task Learning with Document Classification, In EMNLP 2017, 2017.

[3] Mikolov, T., Chen, K., Sutskever, I., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. In NIPS 2013, pp. 3111–3119, 2013.

[4] Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis. In EMNLP 2004, pp. 230–237, 2004.