5N-7　営業日報を対象としたテキストマイニングのための知識辞書の構築*

市村 由美†，中山 康子†，赤羽 俊男‡，三好 みよ子‡，関口 寿一§，藤原 庸祐§
†(株) 東芝 研究開発センター，‡(株) 東芝 情報・社会システム社，§ライオン (株)

1　はじめに

電子化された文書はますます増加しているが、膨大な文書の中から欲しい情報を探したり、文書の集合を分析して傾向を掴んだりするための情報アクセス手段はまだ確立されていない。しかし、ナレッジマネジメントに対する世の中の関心の高さを反映して、営業日報やコールセンターへの問い合わせなどの大量の文書データを分析して内容を瞬時に把握したいというニーズが高まりつつあり、そのための自然言語処理技術としてテキストマイニングと呼ばれる分野が注目されている [1][2]。

我々はテキストマイニングの実践例として、営業日報から成功事例と機会損失事例を抽出する方式を開発し、日報分析システムを試作した [3]。本稿では、成功事例と機会損失事例を抽出するための知識辞書の構築について述べる。

2　知識辞書の構築

営業日報からの成功事例／機会損失事例のひとつとして、次のようなことがらに関する記述と、なぜそうなったのかという要因の抽出を考える。

- 評判が良い、売行きが良い、商談がまとまった
- 評判が悪い、売行きが悪い、商談がまとまらない
- 要望や問題点の指摘

通常このような情報を抽出する手段としては、構文解析を行うことが考えられる。しかし、ここで分析対象としている営業日報は、箇条書きやメモ書きを多用しており、句と句や文と文の関係を構文的に解釈するのは難しい。

そこで、表1に示すようなカテゴリとクラスに分けて、抽出したいことがらに関する知識を記述しておく。表2に知識辞書の記述例を示す。情報抽出の際には、

表 1: 知識辞書のカテゴリとクラス

カテゴリ	クラス
S	売場、社会、店舗、販促 企画、需要、価格、物流、在庫
V	評判、売上、商談、要望

表 2: 知識辞書の記述例

カテゴリ	クラス	キー概念
S	販促	景品を付ける
S	販促	POP を付ける
V	売上	売行きが良い
V	売上	売行きが悪い

カテゴリSとカテゴリVのキー概念同士を組み合わせて検索を行う。

カテゴリSとVは次のような関係にある。

- カテゴリSに属す概念は要因に関する情報、カテゴリVに属す概念は結果に関する情報を表す。
- カテゴリSに属す概念とカテゴリVに属す概念は、格助詞「の」で結びつけられるような連体修飾関係にある。

したがって、この2つのカテゴリの情報を組み合わせることにより、たとえば、「売行きが悪い」に関する情報を、なぜ売れないかという要因や売行きを伸ばすためにとった対策ごとに分類したり、「要望」に関する情報を、何に関する要望や問題点なのか分類することが可能になる。

3　評価実験

3.1　評価データ

提案した方式により構築した知識辞書の有効性を確認するために、ライオン営業日報100枚 (1,231記事、約60,000文字) [1]を用いて、情報抽出の実験を行った。この営業日報は、1人の担当者の1日分の業務記録が

*Architecture of Knowledge Dictionary for Text Mining on Salesperson's Daily Reports

†Corporate Research & Development Center, TOSHIBA Corp., 1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki 212-8582, Japan. E-mail :yumi.ichimura@toshiba.co.jp

[1]この評価データは結果公開の都合上、固有名詞などを一部変更してある。

表 3: キー概念の抽出件数

	概念の種類	記事数	割合 (%)
(a)	カテゴリ S の キー概念のみを抽出	461	37
(b)	カテゴリ V の キー概念のみを抽出	278	23
(c)	カテゴリ S と V の キー概念の両方を抽出	226	18
(d)	どのカテゴリの キー概念も抽出しない	266	22
	計	1231	100

表 4: 関係の抽出件数

	概念間の関係	記事数	割合 (%)
(c-1)	因果関係	194	86
(c-2)	連体修飾関係	11	5
(c-3)	関係なし	21	9
	計	226	100

日報 1〜2 枚に記述されており、1 枚の日報に複数の記事が記述されている。

抽出に用いる知識辞書は、ライオン営業日報 1,263 枚 (6,009 記事、約 394,000 文字) から構築した。これは評価文書には含まれない。また、カテゴリ S に分類されるキー概念数は 245、カテゴリ V に分類されるキー概念数は 54 である。

3.2 結果と考察

表 3 に、キー概念の抽出件数を示す。18%の記事から、2 つのカテゴリに属すキー概念を抽出できた。表 4 に、カテゴリ S と V のキー概念の両方を抽出した記事 226 件 (表 3(c)) について、異なるカテゴリに属すキー概念の組み合わせが因果関係や連体修飾関係を表す記事数を示す。因果関係を抽出できたものは 86%、連体修飾関係を抽出できたものは 5%で、91%の精度で正しく関係を抽出できた。本方式は箇条書きやメモ書きを多用する文書には有効であることがわかった。

正しく関係を抽出できた例 (1-3) と、誤って関係を抽出した例 (4) を示す。

(例 1) 値段的にそんなに安くないので、あまり動きよくない。

(例 2) サンプルを貼付して以来、訪店するたびに売れるようになった。

(例 3) サンプルラッピング分、売れている。

(例 4) 商品 A →サンプルラッピング。商品 B →この 1 ヵ月ほど動きなし。

表 5: キー概念を抽出できなかった記事の分析

	概念の種類	記事数	割合 (%)
(d-1)	カテゴリ S の キー概念のみを含む	49	18
(d-2)	カテゴリ V の キー概念のみを含む	35	13
(d-3)	カテゴリ S と V の キー概念の両方を含む	1	0.4
(d-4)	どのカテゴリの キー概念も含まない	181	68
	計	266	100

例 1 は「安くないので売れない」という価格と売上の因果関係、例 2 と例 3 は「サンプルを付けたので売れた」という販促と売上の因果関係を抽出できた。一方、例 4 は「サンプルを付けたが売れない」という販促と売上の因果関係を誤って抽出した。1 つの記事中に 2 つの製品に関する記述があるためである。今回の実験ではこのような例は少なかった。この営業日報は製品ごとに記事を分けて書くようになっているので、それが成功の要因と言えるだろう。

また、表 5 に、どのカテゴリのキー概念も抽出しない記事 266 件 (表 3(d)) について分析した内訳を示す。キー概念を 1 つも抽出できなかった記事のうち、68%は抽出すべき情報が元々記述されていなかったものであり、残り 32% は知識辞書への登録洩れにより、本来抽出すべき情報が抽出できなかったものである。

4 おわりに

営業日報から成功事例と機会損失事例を抽出するための知識辞書の構築について述べた。本方式によると、メモ書きや箇条書きのような記述を多く含むために構文解析の適用が困難な営業日報からも、因果関係や連体修飾関係を含む情報を疑似的に抽出できる。

今後は、本方式の他の事例への適用可能性を探るとともに、知識辞書の構築の半自動化を目指す。

参考文献

[1] 諸橋, 那須川, 長野. “テキストマイニング：膨大な文書データからの知識獲得 — 意図の認識 —”. 情報処理学会 第 57 回全国大会 5K-03, 1998.

[2] 渡部, 三末. “単語の連想関係によるテキストマイニング”. 情報処理学会 研究報告 FI-55-8 DD-19-8, pp.57-64, 1999.

[3] 市村, 中山, 赤羽, 三好, 関口, 藤原. “営業日報を対象としたテキストマイニング — 成功事例および機会損失情報の抽出 —”. 人工知能学会 第 14 回全国大会 26-06, 2000.