論文の内容の要旨

論文題目

From Opinion to Knowledge
– Extraction of Sentiments and Demands –
意見から知識への転換 – 評価と要望の抽出 –

氏名

Hiroshi Kanayama
金山　博

様々な種類の電子化されたテキスト文書の中では、製品・サービス・企業などに対する評価や要望といった人々の意見が述べられている。これらは企業・個人消費者の双方にとって貴重な情報源となるため、大量の文書から知見を得るために、意見の内容を構造化された情報へと自動的に変換する手法が求められている。本論文では、人々の意見を整理するために好適な意味構造を定義するとともに、意見の内容を高精度で同定するための方法論、さらに特定の分野に適応させるために語彙や構文の構造に関する知識を自動獲得する機械学習の手法について論じる。これらの技法は、実世界の人々の意見を産業界で有効に使える知識へと迅速に効率よく変換するための革新的技術と捉えることができる。

本論文の中核をなす考え方は「節単位の評価表現抽出」と呼ぶ、文書単位・文字単位よりも細かい分析を行い、単語や句の単位の抽出よりも詳細な構造を出力する処理である。具体的には、評判等が書かれた文書から知識を得るために設計された意味構造である「評価フレーム」を出力する。(i)評価表現を正確に検出すること、(ii)同様ないし類似の意味を持つ評価を同一視すること、という評価表現の検出における二つの重要な要求を満たすために、本論文では「木構造変換モデル」を提案する。これはトランスファー方式の機械翻訳で用いられた構文や意味に関する変換操作を模倣するものである。これにより、部分構

文木の結合、動詞の格フレーム解析、語義の曖昧性解消といった機械翻訳を目的として培われた技術が再利用できることとなり、意見の分析に有用な情報を持つ意味構造を高い精度で出力するシステムを、見通しよく、かつ低い開発コストで構築することが可能となる。

本論文の第二の主題は、節単位の評価表現抽出のための語彙を教師無し学習によって獲得する手法である。これは、特定の分野のコーパスの中から、その分野に特化した評価表現の語彙知識を自動的に得るものである。ここで用いる辞書の項目は「極性単位」と呼ばれ、節の極性を決定するための人間が理解できる最小単位の構文構造である。語彙獲得の手がかりとして、「文脈一貫性」、すなわち同じ評価極性が連続して現れやすいという性質を用いて、分野非依存の評価表現をもとに、新たな極性単位の候補を取り出す。そして、コーパス全体における評価表現の密度と極性の一致度の指標を用いて、極性単位の候補から適切なものを、統計的検定により選択する。この結果、専門家にしかわからないような分野に特化した語彙を追加することができるうえ、消費者による評価の具体的内容となるような、製品やサービスに対する新たな良い点・悪い点に関する知識を得ることができるようになる。また本手法は、人手による閾値の設定などが不要であるため、教師無し学習のプロセス全体を全自動で実行できるという利点を持つ。

本論文ではさらに、評判分析の概念を拡張して、人々が製品やサービスに対して求めているものを把握する「要望分析」という新たな課題に取り組む。要望を表す表現には様々なタイプが存在するが、ここでは書き手の要望を表す体言句である「要望対象」の同定に着目する。より多くの要望対象を検出するためには、分野や文書の性質によって異なる書き方の違いを吸収する必要がある。そこで、コーパス全体で述べられている要望の内容の中には共通するものがあるという仮定を用いて、要望を表す語句を得るための新たな構文パターンをコーパスから教師無し学習により獲得する手法を提案する。

論文全体を通じ、現実のビジネスを意識して、実世界のデータを有効に活用するための課題を設定するよう努めている。そして、提案するシステムや構築する言語資源が、複数のアプリケーションや他のコンポーネントによって活用できるような、意味処理の基盤となるように設計する。