TSUKUBAコーパスの様々な分野における有用性の検証

佐藤志織　内山俊郎（北海道情報大学）

◆**概要**：本研究の目的は，製品やサービスに対する書き込み**文書の感情を自動的に分析する（肯定／否定の極性判定など）技術の確立**である．この分野の研究をする上で**TSUKUBAコーパス（楽天㈱がNIIと協力して提供）**は，貴重な研究資源である．しかし，**楽天トラベルという特定の分野**に閉じている．このコーパスで学習した分類器を，**不満調査データセット（Insight Tech社）に対して適用**し，様々な分野の文書に対する有用性を検証した．

実験１（分類器→不満データセット．精度検証）

目的：TSUKUBAコーパスの有用性の検証

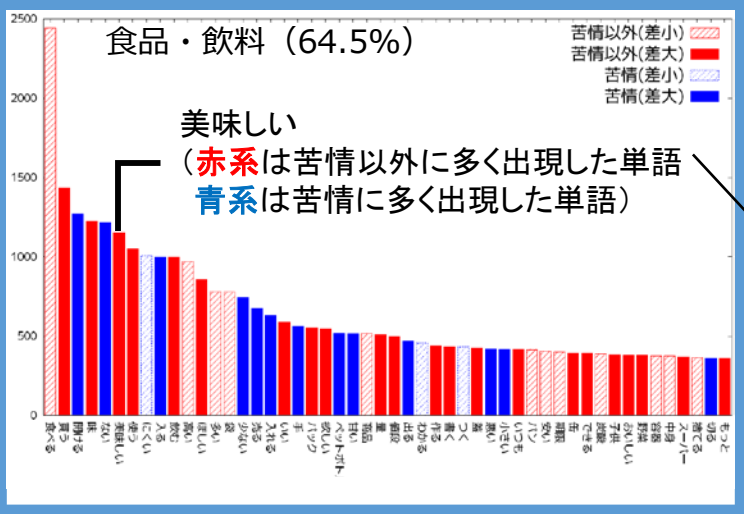
TSUKUBAコーパスは，文章ごとに６種の感情カテゴリが付与されている．このうち３つ（褒め，苦情，要求）が付与されている文章を使い，**ナイーブベイズ（NB)分類器を構成**した．

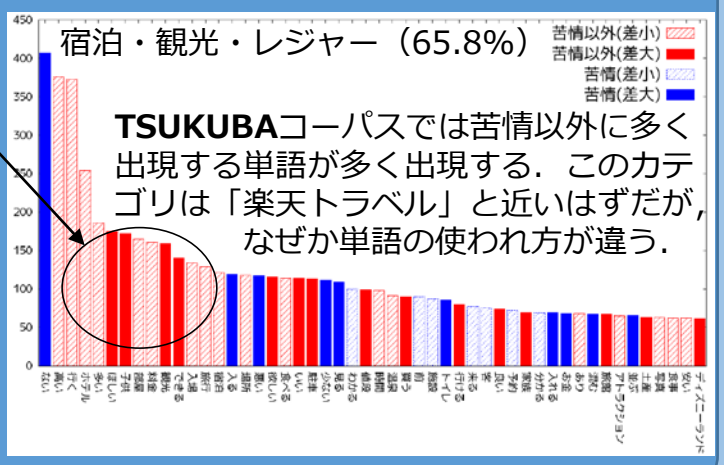
このNB分類器により，不満データセットが**正しく「苦情」に分類できるか検証した．**
精度が低いカテゴリがいくつかありこれがTSUKUBAコーパスの限界と考える．

分野カテゴリ	精度
TSUKUBAコーパス自身	90.6%
自動車	80.8%
業界・業種	79%
・・・中略・・・	
教育	65.9%
宿泊・観光・レジャー	65.8%
仕事	65.6%
食品・飲料	64.5%

[原因1]（TSUKUBAコーパスに無い単語が頻出する）（分類に使われてない）
「仕事」65.6%では、コーパスでは出現しない「給与，働く，残業」が頻出．
[原因2]（TSUKUBAコーパスと使われ方が異なる単語がある）
「食品・飲料」64.5%　（TSUKUBAコーパスとは使われ方が異なる単語がある）
コーパスでは「褒め」に出現する「美味しい」が不満の表現で使われる．

例:「美味しい」による不満表現
・**美味しい** けど値段が高い
・TVで凄く**美味しい**と言っていた食べ物を取り寄せても普通の味だったりする
・肉が柔らか過ぎてブヨブヨ，もう少し歯ごたえがあった方が**美味しい**と思います





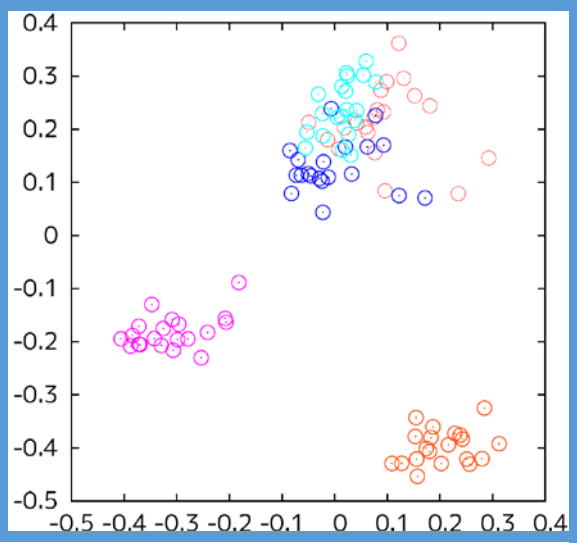
「宿泊・観光・レジャー」65.8%も、使われ方の違いのために精度が低下している．

実験２（可視化）

目的：TSUKUBAコーパスの**感情カテゴリ（褒め，苦情，要求）間の違い**と，**分野カテゴリ（自動車，食品・飲料，楽天トラベル）の違い**は，どちらが大きいのかを，可視化により確認する．

手法：各カテゴリに属する文書から代表的な特徴を情報理論的クラスタリングにより抽出。代表特徴間の距離をJSダイバージェンスにより定義．多次元尺度法により可視化．右図

分野カテゴリの違いが支配的
∴他分野の分析は，難しくなる



結　論

分類実験よりTSUKUBAコーパスの限界を明らかにした．
精度低下の原因は，
・コーパスに存在しない単語
・使われ方（例：美味しい）の違い

可視化の実験から，分野カテゴリの差の方が，感情カテゴリの差よりも大きい．→楽天トラベル．不十分

◆高精度な感情分析のために
・分野特有の単語の使われ方の調査
・様々な分野を網羅した収集