市況の観察に基づく経済動向を対象にした文生成

理学専攻・情報科学コース
1640633
青木花純

1 はじめに

近年, センサーの発達等により観測される時系列データを様々な用途で利用する場面が増えているが，内容理解には専門家による細やかな解釈が必要であり，人の理解を助ける動向概要を示した簡潔な要約文の自動生成技術への関心が高まっている [1].

本研究では，経済指標の時系列データを用いて，動向概要を示す市況コメントを自動生成することを目的とし，過去に観測された時系列数値データとその動向概要を示すテキスト内容の対応関係を学習することによって，新たに観測された時系列数値データの動向概要を示すテキストを自動生成する手法を提案する.

2 提案手法

本研究では，動向概要を示すテキストの自動生成に向けて，以下の手法を提案する.

2.1 概要

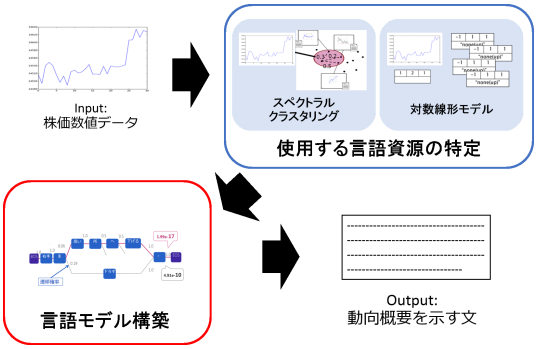


図 1: 概要図

図 1 に概要図を示す．この手法では，新たに観測された時系列データを基に特定した言語資源を基に言語モデルを構築し，確率的に尤もらしいテキストを自動生成する．言語資源の特定には，以下の「対数線形モデルによる識別器」と「時系列データの類似度に基づくクラスタリング」を用いた.

2.2 言語資源の特定

2.2.1 対数線形モデルによる識別器

まず，過去に観測された時系列データを Symbolic aggregate approximation を用いて次元圧縮を行い，表 1 に示す動向内容を示すラベル (以下，中間表現と呼ぶ) を人手で付与する．その後，対数線形モデルを用いて式 1 に示す中間表現の識別器を構築する.

$$P(r|d) = \frac{1}{Z_{d,\boldsymbol{w}}} exp(\boldsymbol{w} \cdot \phi(d, r)) \quad (1)$$

この識別器を用いて，新たに観測されたデータの中間表現を判別し, 同じ中間表現が付与された過去のデータと対で収集されたテキストを言語資源として特定する.

表 1: 中間表現

中間表現	動向内容
UU	上昇し続けた
SS	下落し続けた
DD	あまり変化はなかった
US	上昇したが，その後変化がなくなった
SD	あまり変化がなかったが，その後下落した
DU	下落したが，その後上昇した
UD	上昇したが，その後下落した
SU	あまり変化がなかったが，その後上昇した
DS	下落したが，その後変化がなくなった

2.2.2 時系列データの類似度によるクラスタリング

この手法では，新たなデータが観測された際に過去のデータと共にクラスタリングを行い，新たに観測されたデータと同じクラスタに分類された時系列データと対で収集したテキストを言語資源として特定する．その際，各言語資源に対して新たに観測された時系列データの類似度により重みづけを行なう.

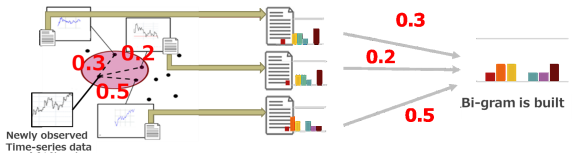


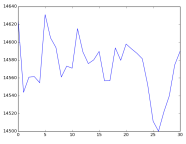
図 2: クラスタリングによる言語資源の特定

クラスタリングにはスペクトラルクラスタリングを用いた．この手法はデータをノード，データ間の類似度をノードの距離として無向グラフを作成し, 式 2 に示される Normalized Cut(NCut) 関数を用いて，クラスタ内の結合を考慮したグラフカットを行うことでデータをクラスタリングする手法である.

$$\begin{cases} NCut(C_1, C_2) = \frac{cut(C_1,C_2)}{assoc(C_1,V)} + \frac{cut(C_1,C_2)}{assoc(C_2,V)} \\ assoc(C_1, V) = \Sigma_{u \in C_1, t \in V} w(u.t) \end{cases}$$

本研究では時系列データの類似度として, 時系列データの各点の距離を総当たりで計算し，作成した距離行列を基に計算する Dynamic Time Warping(DTW) 距離 [3] および DTW 距離に制約パラメーター r を導入することによって, 時系列データの接頭/接尾で発生する類似度の誤差をなくした Prefix and Suffix-Invariance DTW(ψ-DTW)[4] を用いた．これらの評価指標は，共通してデータ中の挙動の周期性等が違う場合でも適切に類似度を示す事が出来るが，短所として計算量が多いことがあげられる.

表 2: 生成されたテキストの一例

時系列数値データ	正解文	
	一時上げ幅を拡大した。	
手法	ラベルアルゴリズム	生成文
識別器	識別結果：「UU」	上げ幅, を, 拡大, し, た,。 ⋯, EOS
クラスタリング	DTW	上げ幅, を, 拡大, し, た,。 ⋯, EOS
	ψ- DTW	一時, 下げ, 幅, を, 拡大, し, た,。 ⋯ EOS

2.3 言語モデルによる文生成

特定された言語資源を基にバイグラムを用いた言語モデルを構築し，動的計画法を用いることによって確率的に尤もらしい単語の組合せを獲得し，文を生成する．ngram によるモデルの特徴として，文長が長い文ほど尤度が低くなってしまう傾向があるため，使用する言語資源は言語資源中の最大文長に合わせて，仮想の単語として番号付きの null ラベルを挿入した．

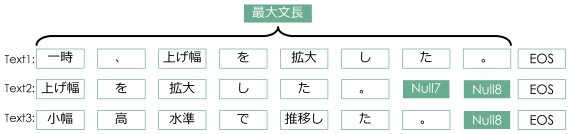


図 3: 仮想の単語 null の挿入

3 実験

本章では提案手法を用いて，新たに経済指標の時系列数値データが与えられた際に，データの動向概要を示すテキスト生成を行い，評価を行う．

3.1 実験設定

使用する時系列データおよび対となる言語資源は前場と後場の各時間帯に分けて収集した．期間は 2013 年 2 月 25 日〜2014 年 12 月 30 日，451 日分の計 902 個である．また，2 つの特定手法の比較を行うため，対数線形モデルにおける中間表現は「Up/Down/Stable」とその組み合わせ (例「UpUp」) とし，SPC におけるクラスタ数を 3 および 9 と設定した．

3.2 生成結果および考察

中間表現の識別精度は 10 分割交差検定を行った結果，要素数 3 つの際には約 47%，9 つの際には約 26%であった．

生成テキストの例を入力とした時系列データと正解テキストと共に表 2 に示す．両手法とも生成されたテキストには短文が多く見られた．その理由として，訓練データの多くが短文であったことが考えられる．また、意味的に冗長なテキストが生成された例も見られたため，形容詞や動詞などの重要単語に文法を考慮したモデルを構築することが必要であると考えられる．

また中間表現付与のコストがかかる識別器による手法と比較して，クラスタリングによる手法は手法適用のコストが低いと考えられる．しかし，クラスタ数の適切な設定が必要であるという点では．中間表現の設定と同様である．このように中間表現の設計や計算量を考慮すると，両手法ともに複雑な事象を述べるようなテキスト生成には向いていないと考えられる．

また，生成されたテキストの例のように，識別器による生成テキストと比較して，クラスタリングを用いる手法による生成テキストの方が，テキスト中の単語の多様性が高い傾向があった．これは中間表現の設定による制限のためであると考えられ，適切な中間表現を設定する事の難しさを示している．一方で，クラスタリングを用いる手法も現段階ではクラスタ数を設定しているが，今後はクラスタ数を自動推定するノンパラメトリックな手法の適用が必要であると考えられる．

4 おわりに

本論文では経済指標を用いて，時系列データの動向概要を示す市況テキストの自動生成に取り組み，言語資源を特定し言語モデルを構築する手法を提案した．実験結果やコストの観点から，クラスタリングによる言語資源の特定手法が優位であると考えているが，適切なクラスタ数の決定手法が必要であると考えている．今後はクラスタの自動推定などに深層学習を用いるなどして，人手による制限をなくした手法の提案，実装および比較評価を行う予定である．

参考文献

[1] Soichiro Murakami, Akihiko Watanabe, Akira Miyazawa, Keiichi Goshima, Toshiki Yanase, Hiroya Takamura, Yusuke Miyao, “Learning to Generate Market Comments from Stock Prices”, 55th Annual Meeting of Association for Computational Linguistics(ACL), 2017

[2] Lessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu,”A Symbolic Representation of Tie Series, with Implications for Streaming Algorithms,”In SIGMOID workshop, 2003

[3] Ding Hui, Trajcevski Goce, Scheuermann Peter, Wang, Xiaoyue, Eamonn Keogh, “Querying and mining of time series data: experimental comparison of representations and distance measures”, Proc. VLDB Endow1 (2): 1542-1552,2008

[4] Silva. D. F.,Batista G. E. A. P. A and Eamonn Keogh, “Prefix and Suffix Invariant Dynamic Time Warping”, IEEE international Conference on Data Mining, 2016