障害情報レポートに対する同時関連文章圧縮

小平 知範　　宮崎 亮輔　　小町 守
首都大学東京
{kodaira-tomonori,miyazaki-ryosuke}@ed.tmu.ac.jp, komachi@tmu.ac.jp

1 はじめに

iPhone が壊れた！ 突然このようなトラブルに見舞われたとき，みなさんはどのようにサポート対応を受けるだろうか．チャットやメールのようにテキストで状況を相談することや，ストアで対面の相談を受けたりすることもできる．しかしながら，テキストベースでは細かいニュアンスを伝えることができず，対面だと直接ストアに足を運ばなければならない．そこで，電話によるサポートが広く用いられている．

企業ではそういったトラブル対応のためにコールセンターを利用している．コールセンターにはその対応をまとめたログが大量に溜まっている．これらのデータは日々膨大な量が蓄積されているが，多くの場合ログデータは保存されていくだけで有効活用されていない．類似した障害に対応するため，蓄積されたレポートを参照する必要があるが，自由に記述されたレポートをそのまま参照するには時間がかかる．そこで，これらのログデータから自動的に重要箇所抽出することで，メンテナンスなどの保全作業の効率化やマニュアル改善に役立てる等の有効活用ができる．

このようなコールセンターのログデータからのテキストマイニングの一つとして，本研究ではシステム障害に対するレポート文章から重要箇所を抽出するというタスク（文章圧縮）に取り組む．重要箇所の抽出は文圧縮技術を用い，重要でない単語を文章中から削除する手法で行う．例えば今回用いるデータの一部を図 1 に示す．このレポートは，システム障害に対する問い合わせに対しての一連の対応について記述されている文書である．システム障害が起こった状況，原因，処置の 3 つの文章で構成されており，それぞれの重要箇所が抽出されている．重要箇所は状況，処置の文章に 1 つずつ，原因の文章は 2 つの観点から重要箇所が抜き出されている．コールセンターに蓄積されているのは状況，原因，処置の文章のみであり，ここから自動的に情報を抽出するというタスクである．

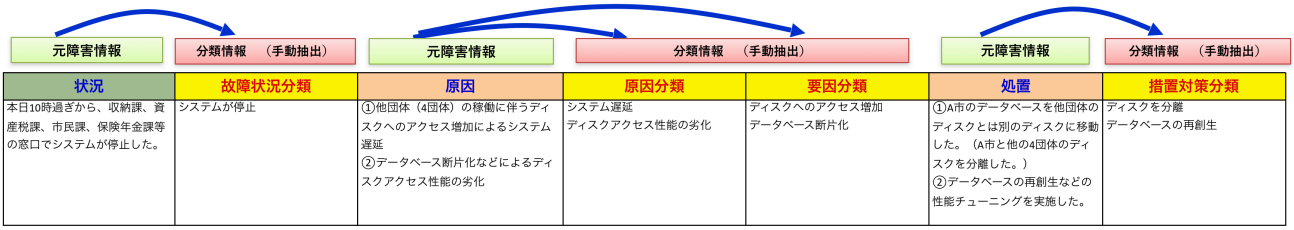
自由記述された文書からの情報抽出としては，先行研究では文を要約する手法 [1, 2] や抽出型文書要約手法 [3, 4] などがある．これらは一文に対し文圧縮や文要約を行うものや，文書から重要文抽出を行っている．しかしながら，これらの手法は本タスクのように 1 つのイベントについて複数の観点から記述した文章が存在する場合に，相互の関係を考慮できない．また，複数文書要約を扱う Wan ら [5] の手法は複数の文書の入力から要約を生成するものがある．しかし，複数文書要約は各文書がそれぞれ違う観点で記述されたものに対して，観点ごとに要約を生成するものではない．

そこで，本研究では，関連する複数の文章を同時に考慮する文章圧縮手法を提案する．マルチタスク学習のように，複数文章で同時に学習させることで共通の情報を獲得し，再現率の高い予測をできるようにする．ニューラルネットを用いた文圧縮の先行研究をベースとし，各文章の情報を考慮したモデルを考案する．

2 関連研究

文圧縮の先行研究には，Encoder-Decoder モデルを用い，文中の単語に対してラベル付けをする系列ラベリング問題として文圧縮を解く Filippova ら [2] の手法がある．彼女らは入力文中から文圧縮後に含まれない単語には 0 ラベル，含まれる単語には 1 ラベルを与えて学習することで文圧縮を行っている．本研究では Filippova ら [2] と同様に系列ラベリング問題として重要箇所抽出を行うが，入力は文ではなく文章である．

ニューラルネットを用いたマルチタスク学習には，単言語から複数言語に翻訳する Dong ら [6] の手法がある．彼らは単言語から多言語への翻訳をするために，原言語側の単一のエンコーダに対し，デコーダを各言語ごとに用意している．本研究で扱うタスクは機械翻訳ではないが，同じ文書内から抽出された複数文章を扱うため，相互に情報を共有する機構を追加する．従って，彼らとは異なり，エンコーダを共有するので



状況	故障状況分類	原因	原因分類	要因分類	処置	措置対策分類
本日10時過ぎから、収納課、資産税課、市民課、保険年金課等の窓口でシステムが停止した。	システムが停止	①他団体（4団体）の稼働に伴うディスクへのアクセス増加によるシステム遅延 ②データベース断片化などによるディスクアクセス性能の劣化	システム遅延 ディスクアクセス性能の劣化	ディスクへのアクセス増加 データベース断片化	①A市のデータベースを他団体のディスクとは別のディスクに移動した。（A市と他の4団体のディスクを分離した。） ②データベースの再創生などの性能チューニングを実施した。	ディスクを分離 データベースの再創生

図 1: 障害レポートに対するアノテーション例

はなくエンコーダの隠れ層を共有する機構を用いる．

コールセンターの障害レポートログは箇条書きや読点がなく‘(1)’のような表現が文頭と文中に混在するものが含まれるため，文分割を前提とする手法は適用するのが困難である．本研究では Filippova ら [2] の手法をベースとし，文章を圧縮する形でアプローチする．今回のタスクは 1 文書中から各クラスに分類された関連文章データを用いるため，各クラスの情報を共有するマルチタスク構造のモデルを提案する．

3 障害レポートの同時関連文章圧縮

3.1 タスク設定

本研究で用いるデータには，1 件のシステム障害対応に対して，状況，原因，処置の 3 つの文章がある．重要箇所のアノテーション例を図 1 に示す重要箇所は状況，原因，処置の文章からそれぞれ故障状況分類，原因分類・要因分類，措置対策分類 4 つについて抜き出されている．

この例では各文章内の各文に重要箇所が一つずつあるが，重要箇所が存在しない文を含むこともある．また，文章からは重要箇所が抜き出せない場合もある．原因の文章からは原因分類・要因分類の 2 つの観点から重要箇所が存在する．そのため，原因分類と要因分類には重複が存在する場合がある．

本研究では，文章から重要箇所の抽出を行うために文圧縮手法を適応し，文章中の重要箇所に含まれる単語には 1 を，含まれない単語には 0 のラベルを振る系列ラベリング問題として解く．今回扱うデータは重要箇所をより多く示すことが重要であるため，再現率の高いモデルを作成する必要がある．

3.2 手法

3.2.1 LSTM を用いた文圧縮

Filippova ら [2] は，文圧縮タスクを文中の各単語に対して圧縮後の文に入るか否かを推定する系列ラベリング問題として解いている．彼女らの手法は Long Short Term Memory (LSTM) による sequence to sequence を用いている．

LSTM を用いた系列ラベリングを解くネットワークを図 2 に示す．エンコーダ側に入力文を後ろの単語から順に 1 単語ずつ単語の埋め込みベクトルを入力していく．入力文の先頭に到達したら，GO シンボルを入力し，デコードを開始する．デコーダ側では入力文を先頭の単語から順に 1 単語ずつ入力し，ラベルを出力する．このラベルは 3 種類で，圧縮文として出力する単語ならば 1，削除するならば 0，GO と end-of-sentence のシンボルを入力した際は EOS としている．

単語の埋め込みベクトルには SkipGram モデル [7] で学習済みの分散表現を用いている．また，ラベルを予測する際に単語の分散表現に加えて，前単語のラベル情報，親単語のラベル情報，親単語の分散表現を連結して入力している．前単語のラベル情報は，その単語がどのラベルを振られたかの 3 次元の One-Hot ベクトル，親単語のラベル情報は，振られたラベルが圧縮文に含まれるか否か，まだラベルが振られていないかの 3 次元の One-Hot ベクトルとしている．親単語の情報は，入力文を依存構造解析した結果から得る．

ネットワークの構造は，3 層の LSTM を重ねており，最終層以外の LSTM 層に dropout を適用している．最終層の LSTM の出力に softmax をかけ，ラベルを推定する．

3.2.2 提案手法: 関連文章を考慮した文圧縮

本研究では Filippova らの文圧縮の手法をもとに，複数文章を互いに考慮し，同時に圧縮を行う（図 3）．

状況，処置の文章に 1 つずつ，原因の文章に対しては重要箇所抽出の種類が 2 つあるため 2 つ，の合計 4 つのエンコーダを用意する．同様にそれぞれに対応するデコーダも 4 つ用意する．デコーダ側にエンコーダ側の隠れ層を渡す際に，エンコーダの 4 つの隠れ層を連結し，次元数を 1/4 にする線形変換を行う．この線形変換は各デコーダごとに異なった変換を行う．こうして，複数文章中の情報を一度統合することによって，

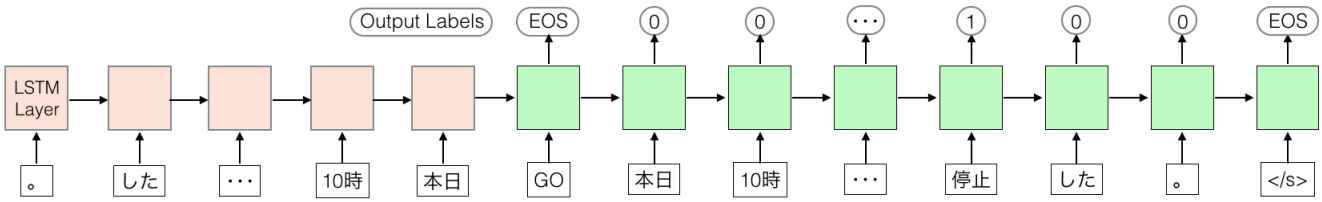


図 2: LSTM を用いた文圧縮を系列ラベリングとして解くネットワーク

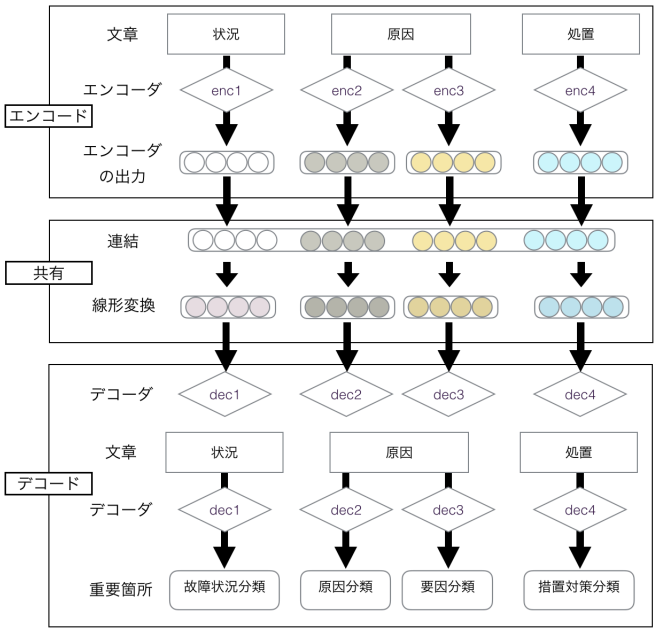


図 3: 文章間の情報を考慮した同時関連文章圧縮モデル

複数の文章全体を考慮して重要箇所を抜き出すことができる表現を作成する．

各ネットワークの構造は図 2 に示したものだが，GO ラベル入力の前に 4 つの文章で用いたエンコーダの隠れ層を連結し，線形変換を行い，デコーダへと渡す．

4 実験

4.1 データ

本研究では，富士電機 (株) でのシステム障害の 511 件のレポート文書に対して 3.1 節のアノテーションを付与したものを使用する．今回用いるデータ内には 2 種類のシステムについての障害レポートがあり，それぞれのレポート文書は 269 件と 242 件である．

重要箇所のアノテーションは 4 人が行ったが，そのうちの 1 人のデータを正解データとして扱う [1]．4 人の重要箇所アノテーションの各一致率を表 1 に示す．一致率の計算には Jaccard 係数をペアワイズで計算したものの平均を用いた．F_2 スコアは正解データと他 3 人のアノテーションを評価した平均を表している．

[1] アノテーションがマークアップではなく記述方式であるため，誤字が含まれることがあった．そのため，正解データには最もそういった間違いの少ない人のデータとした，

	故障状況	原因	要因	措置対策
Jaccard 係数	0.578	0.662	0.520	0.704
F_2 スコア	0.711	0.694	0.562	0.806

表 1: 重要箇所抽出アノテーションの一致率

4.2 実験設定

比較手法として以下の 2 つの手法を用いる．

- Filippova ら [2] の手法 (SingleLSTM)
- SingleLSTM のエンコーダの 4 つの出力を連結し，線形変換をし，デコーダへ渡す手法 (MultiLSTM)

各文章に対し，形態素解析器 MeCab (Version: 0.996, ipadic) を用いて形態素解析を行う．また，文章中に含まれる記号列（ディレクトリやコマンド等）は削除している．時間表現は ‘TIME’，①や(1)などの数字表現は ‘ENUM’，トレーニングデータ内で頻度 8 以下の英数字列は ‘ALPHA’ というシンボルに変換している．今回使用するデータは文分割が明確に行われておらず，正確に分割を行うのが困難であるため，入力は文章であり，文分割は行っていない．

実験の際は，このデータのうち 311 件をトレーニングデータ，100 件を開発データ，100 件をテストデータとして，5 分割交差検定を行う．評価方法は文章中の重要箇所を選んだものを正解とし，評価尺度として F_β を用いる．このタスクではユーザーに重要箇所をなるべく多く提示することが重要なので，式 1 の β の値を 2 として F_2 スコアを計算し評価する．

$$F_\beta = (1-\beta^2)\cdot\frac{precision\cdot recall}{(\beta^2\cdot precision)+recall} \quad (1)$$

4.3 パラメータ設定

単語の埋め込みベクトルの次元数は 256 次元，LSTM の隠れ層の次元数は 256 次元．先行研究とは異なり，単語の埋め込みベクトルと LSTM の隠れ層について

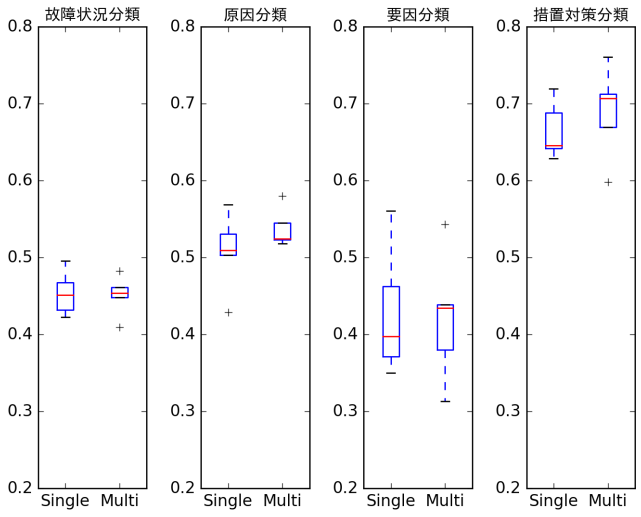


図 4: 各文章ごとの実験結果（F_2 スコア）

はランダムに初期化し学習する．本研究では親単語の情報は用いていない．エンコーダとデコーダ共に，LSTM を 3 層に重ねている．dropout の確率を 0.2，最適化は AdaGrad を用い，学習率を 0.01 に設定する．エポック数 50 回で学習し，開発データで F_2 スコア (式 1) が最大となったエポックのモデルを最終的なモデルとして選ぶ．

4.4 結果

5 分割交差検定の結果を図 4 に示す．措置対策分類の予測結果は他よりスコアが高く，他分類よりアノテーションの一致率と近い値が出ている．

MultiLSTM で他文章の情報を共有することで，精度の向上が見られた．要因分類では，分散が他文章より大きくなっている．

5 考察

今回はエンコーダの各出力を合わせ，線形変換を行いデコーダへと渡す簡単な機構を用い，多くのタスクで精度が向上した．

実験の結果，措置対策分類以外では，表 1 で示した人手の F_2 スコア（上限値）より 0.20 ポイント以上低い結果が出ている．今回学習データ量が少ないため，分類できないデータが多いと考えられる．または，用いた手法が文を入力として想定しているため，文章のような長い単語列に対応できなかったものだと考えられる．

一方，MultiLSTM を用いて他文章の情報を共有することで単文章では得られない情報を獲得できたため，精度が向上したと考えられる．要因分類では，2 手法ともに分散が高く，アノテーションの一致率同様に精度も低い．要因分類の重要箇所は原因分類の重要箇所に強く依存するため，今回用いた線形変換のような簡単な機構では捉えきれなかった．故障状況分類の重要箇所にも他分類の重要箇所に関連する部分は少なからずあるため，他分類での重要箇所抽出結果を考慮する機構を導入する必要がある．

6 おわりに

本研究ではニューラルネットに基づく文圧縮を用い，関連のある複数の文章から同時に重要箇所抽出を行った．各文章で用いたエンコーダの隠れ層と連結させてから，デコーダへと渡すことで相互に情報を共有できるような機構を提案した．

今回実験を行った手法では，人手によるアノテーションの一致率よりかなり低い値が出た．学習には少量のデータしか用いていないため，入力単語の情報を豊富にする必要がある．また，文単位の重要箇所抽出手法をもとにしているため，文章という長い単語列を扱うための機構が必要である．

謝辞

本研究では，株式会社富士電機様よりデータの提供と重要箇所抽出のアノテーションをしていただきました．ここに感謝の意を示します．

参考文献

[1] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *EMNLP*, pp. 379–389, 2015.

[2] Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. Sentence compression by deletion with LSTMs. In *EMNLP*, pp. 360–368, 2015.

[3] Yuta Kikuchi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. Single document summarization based on nested tree structure. In *ACL*, pp. 315–320, 2014.

[4] Chen Li, Xian Qian, and Yang Liu. Using supervised bigram-based ILP for extractive summarization. In *ACL*, pp. 1004–1013, 2013.

[5] Xiaojun Wan and Jianguo Xiao. Graph-based multimodality learning for topic-focused multi-document summarization. In *IJCAI*, pp. 1586–1591, 2009.

[6] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. Multi-task learning for multiple language translation. In *ACL-IJCNLP*, pp. 1723–1732, 2015.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pp. 3111–3119, 2013.