宇宙機関連異種テキストデータからの知識発見の試み

森 健史[1]　大久保 梨思子[1,a)]　植田 泰士[1]　片平 真史[1]　天笠 俊之[2]

概要：本稿では，設計過誤の再発防止を目的として，新たに開発を行う宇宙機設計の仕様情報のテキストデータ群に対して， 過去に宇宙機において発生した不具合情報のテキストデータ群の中から強く相関するテキストの組合せを抽出する方法について検討する．宇宙航空研究開発機構（以下，JAXA）では，製品開発の過程で生じる多くのデータが文書等で記録されており，後続の製品開発へ有効に活かすことが期待されている．中でも製品の不具合情報は，製品開発数が限られ，開発ライフサイクルの長い宇宙機開発においては，過去の失敗経験の知識を後続の開発に活かすことができる重要なデータである．その一方で，事象・原因等が記述されている不具合情報は，対象とするシステムが人工衛星，ロケット等多様でありソフトウェアからハードウェアまで対象分野が広範であること，書き手の経験等により文書表現の揺らぎがあるといった特徴がある．また宇宙機設計の仕様情報は，比較的抽象度の高い情報であり，不具合情報に記述されている情報とは記述レベルや粒度等が異なる．このため，両者のマッチングは単純に行うことができない．そこで，本稿では，まず特定製品情報を記述した仕様情報のテキストデータと，不具合情報のテキストデータを組み合わせ，仕様情報にある製品の設計過誤につながる知識を，検索システムを通して発見することを試みた．検索の際，仕様情報に対して不具合情報が有用であったかの人の価値判断を，畳み込みニューラルネットで学習した結果，特定ドメインのデータに対し 50%の適合率で仕様情報に対して有用な不具合情報を引き当てることが出来た．

Towards Knowledge Extraction from Heterogeneous Text Data Related to Space Systems

1. はじめに

宇宙航空研究開発機構（以下，JAXA）では，製品開発の過程で生じる多くのデータが文書等で記録されており，後続の製品開発への有効活用が期待されている．中でも製品の不具合情報は，製品開発数が限られ，開発ライフサイクルの長い JAXA の宇宙機開発においては，過去の失敗経験の知識を後続の開発に有効に活かすことができるデータとして期待されているが，十分な活用には至っていない．

本稿では，過去発生した不具合と同様の設計過誤につながり得る知識を発見することを目標に，新たに開発を行う宇宙機設計の仕様情報のテキストデータ群（以下，仕様情報）に対して，過去に宇宙機開発で発生した不具合情報のテキストデータ群（以下，不具合情報）から相関の強い情報を，検索システムを通して抽出することを試行した．不具合情報と仕様情報のテキストデータは，各々記述内容や抽象度について性質が異なる．例えば，不具合情報は具体的に発生した事象・原因が記述されている一方で，設計情報では比較的抽象度の高い設計に関する記述がされている．そのため，両データを単純に文字列として扱いマッチングさせた場合，期待した対応関係を発見することは難しい（2 章）．

上述のような異種データを組み合わせて知識を発見するためには，性質の異なるデータを関連付けるラベルを付与する等，データに対するアノテーションを行なうアプローチが考えられる．しかし，本研究が対象とするデータは年月とともに傾向，利用目的が変わるため，データ収集時点で必要十分なラベル設計を行うことが難しい．また，一度付与したラベルを修正するコストが高い等の課題がある．

以上を踏まえ，本試行ではデータへのアノテーションを前提とせず，仕様情報に対して関連する不具合情報に，設計過誤につながる知識が含まれているかどうかのラベルを

[1] 国立研究開発法人 宇宙航空研究開発機構
2-1-1 Sengen, Tsukuba, Ibaraki, 305–8505, Japan
[2] 筑波大学 計算科学研究センター
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
[a)] okubo.naoko@jaxa.jp

ⓒ 2019 Information Processing Society of Japan

表 1 仕様情報の特徴

主な記述内容	特定製品の設計根拠 (解析結果等を含む) 機能等の設計結果
データ特徴	開発フェーズによっては抽象度が高い記述
データ量	単語数：約 47 万単語（ある衛星の例）

表 2 不具合情報の特徴

主な記述内容	不具合の事象，原因，対策記録等の事実情報 他製品への水平展開要素等の解釈情報
データ特徴	製品及び企業横断的に収集され，記述者は多数 情報の欠落があるデータも存在
データ量	不具合件数：41,923 件 1 件あたりの平均単語数:約 560 単語

人手で付与し，それを教師データとして機械学習手法を適用することを試みた（3 章）．具体的には，畳み込みニューラルネット（以下，CNN）で，地球周回衛星の仕様情報，過去の不具合情報，両者の組合せの価値判断結果を学習することにより，各テキストデータ群の再整備を行うことなく，データを活用することが可能なことを示した．

2. 宇宙機関連テキストデータの特徴と課題

本稿の知識抽出試行で取り扱う 2 種類のテキストデータの特徴について述べる．1 つ目は，新たに開発を行う宇宙機設計の仕様情報である．これは，宇宙機開発におけるアウトプットの 1 つであり，段階的プロジェクト計画（Phased Project Planning）における技術レビュー等にて参照される．2 つ目は，過去の宇宙機において発生した不具合情報であり，JAXA 内においてウェブシステムで運用されるデータベースである．不具合情報蓄積の主な目的は，各不具合の解決状態管理に加えて，同様の不具合の混入・残留防止を目的としたユーザからの検索，参照である．仕様情報と不具合情報の特徴を，**表 1**，**表 2** に各々示す．

本試行において，仕様情報を手掛かりに，関連する不具合情報を発見することを考える．すなわち，仕様情報を検索キーとした不具合情報の検索と考えることができる．表 2 の通り，不具合情報は，実際に発生した事象やその原因を記述しているため，記述内容は具体的である反面，書き手の経験等による文書表現の揺らぎや情報欠落が見られる．また，製品（ロケット，衛星，地上設備など）や，不具合の原因種別（経年劣化，論理過誤，運用ミスなど）は横断的に収集され，異なるドメインの内容が混在している．従って，表 1 に示されるような，特定の製品情報のみが記述された仕様情報とは性質が異なる．このような異なる特徴を持つ文書の関連付けは，単純なキーワードマッチングで行うことは難しい．

このことを確認するために予備的な検証を行った．単語の出現頻度から文書内の単語の重要度を算出する手法である Okapi BM25[1] を用い，仕様情報と不具合情報から重要度の高い 3 単語を，それぞれのキーワードとして抽出し，

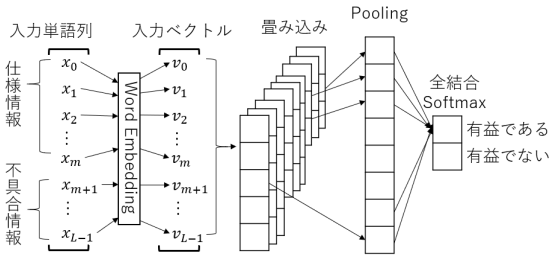

図 1 提案手法

突合せを行った．類似度の上位 20 件を，被験者による有用性判断をしたところ 8.25%と非常に低い適合率となった．

予備検証の中で，キーワードマッチングによる引き当てを困難にしていると考えられる 2 つの事象を識別した．（事象 1）テキストデータに宇宙機開発特有の意味がある単語が含まれる場合に，異なる文脈の不具合が引き当たってしまう事象である．具体的には，「ホイール」という単語は，地上設備の車に装着される車輪を指す場合と，宇宙機の姿勢制御に用いられるリアクションホイールの文脈で使われる場合がある．（事象 2）表記そのものに揺らぎがあるため、同義の単語群が各々別のものとして扱われてしまう場合である．略語などがこのケースであり，例えば，「STT」という単語は Star Tracker の略語として用いられており，記載に略語を用いたか否かは，書き手によって不確定である．

3. CNN を用いた異種文書の関連付け手法

性質の異なる文書間の関連付けを行うため，近年多くの分野において高い精度が確認されている CNN[2] を用いることとした．特に仕様情報と不具合情報の関連付けの判断は複雑であり，高次の非線形回帰が必要であるとの仮定による．**図 1** に提案手法の概要を示す．

3.1 教師データ

CNN の学習に用いる教師データは，仕様情報と不具合情報の組が製品の設計過誤につながる知識として有益か，そうでないかの 2 クラスを被験者が判断した結果とする．被験者は，同程度の業務経験のある 3 名とし，2 名以上が有益であると判断したものを正例とした．

3.2 CNN の構成

CNN への入力データを，仕様情報と不具合情報を単語分割し，それらを連結した単語列とした．文を単語単位に分かち書きするための形態素解析に MeCab[3] を用いた．単語の分散表現への変換は，分散表現学習手法として広く適用されている word2vec[4] を採用した．分散表現の学習データには，不具合情報に含まれる文書と，Wikipedia から宇宙機開発で頻出する用語のトピック文書を用いた．図 1 において入力単語列中の i 番目の単語を x_i，対応する

ⓒ 2019 Information Processing Society of Japan

表 3　word2vec 学習パラメタ

分散表現学習モデル	skip-gram
窓幅	10
ベクトルサイズ	200
エポック数	5
ネガティブサンプリング数	10
学習率	0.025
頻度閾値	10
バッチ数	1000

分散表現を v_i で表現する．分散表現 v_i を N_{word} 次元とし，入力の単語長を L としたとき，入力された単語列は，$\mathbb{R}^{L \times N_{word}}$ の行列となり，これを CNN への入力とした．

教師データを被験者の試行により作成するため，多量の教師データの確保が難しいことを前提とし，モデルはパラメタ数の少ないシンプルな設計とし，ReLU を活性化関数とする畳み込み層，MaxPooling 層，全結合層と Softmax 層から成る (図 1)．人手による教師ラベルの判断は，単語の並びと，それらの並びの出現組み合わせにより表現可能であるという仮定の下で設計を行った．

畳み込み層のフィルタは，複数の窓幅を持ち，窓幅内の単語の並びから判断に寄与する単語列を学習する．入力は単語の分散表現の列であるため行方向へのスライドは不要であり，窓幅を L_w とした場合，フィルタは $\mathbb{R}^{N_{word} \times L_w}$ の行列となる．学習したフィルタを用いて畳み込み操作を行うことで，判断に寄与する単語列の信号が抽出される．

次に，MaxPooling 処理を行うことで，各フィルタ特徴と最も適合している単語列の信号のみが特徴として抽出される．これにより単語列の文書中の位置情報は失われるが，特徴となる単語列の出現位置は書き手の特性により変動するパラメタであり，これは許容されると考えた．

更に全結合層で，抽出された信号を総当たり的に組み合わせ学習を行うことで，文書の特徴を捉え，有益かの判断を行う．Softmax 層は 2 つの信号を出力し，一方を有益である確率の大きさ，他方を有益でない確率の大きさとし，確率の高いクラスを推論結果とした．損失は，分類タスクで一般的に用いられる交差エントロピーとして学習を行う．

本提案手法では，2 つの文書の関連性判断のために明示的に特徴量比較は行っていないが，CNN モデルの学習過程にて関連性を捉えることを期待した．つまり，教師データ中には，同一の仕様文書に対して，有益な不具合文書の組とそうでないもの双方が存在しており，それらに対して最適化を行う中で，両文書の特徴が判断に必要であることを捉えると考えた．

4. 評価

提案手法と被験者間の判断の比較実験を行った．また，2 章で述べた通り，単純なキーワードマッチングによる引き当てが難しいことを提案手法と対比するため，単語の分散表現の評価として，単語の出現頻度を源泉とする Latent Dirichlet Allocation[5]（以下，LDA）を用いた場合の精度と比較した．教師データとして用いる仕様情報は，ある地球周回衛星の仕様文書から技術的な内容を含む 20 文書とした．これらを検索クエリとした 764 組について教師ラベルを用意し，98 組が正例，666 組が負例となった．

表 4　モデルの学習パラメタ

仕様情報長	221
不具合情報長	431
分散表現サイズ	200
窓幅	3，4，5
フィルタ数	128
エポック数	200
バッチ数	30
学習率	0.0001
最適化手法	Adam

表 5　実験結果

		被験者の判断	
		有益でない	有益である
推論結果	有益でない	420（355）	24（43）
	有益である	246（311）	74（55）

4.1 学習パラメタ

3.2 節で述べた word2vec による分散表現学習用データとして，語彙 101,991 を含む 760,337 文書を用意した．その他の実験条件を**表 3** に示す．CNN の学習パラメタを**表 4** に示す．データ不均衡による学習結果の偏りを避けるため，同数の正例と負例をサンプリングして学習を行った．

4.2 実験結果

教師データに対して leave-one-out 交差検証を行った結果を**表 5** に混同行列としてまとめる．括弧内は単語の分散表現に LDA を用いた際の結果である．提案手法は，精度 64.7%（適合率 23.1%，再現率 75.5%）となり．仕様文書ごとの精度は，平均値/中央値/最大値/最小値が，それぞれ 64.2%/67.5%/85.0%/42.5%であった．

検索のユースケースにおける精度である，適合率，再現率について述べる．CNN は Softmax 層の出力を判断結果の確率値として扱うことができる．各仕様文書の判定結果を，確率値順にソートした上位 5 件に対する Mean Average Precision（以下，MAP）は 66.4%，確率値が 0.9 以上の値で正例と判断しているものだけをユーザに提示する場合，適合率は 50.0%となった．再現率については，不具合文書数が 41,923 件あり，それらすべてを人間が判断することは，リソース制約上不可能であり計測ができていない．しかし，検索結果の信頼性を見る上で重要な指標であり，今後，擬似的にでも確認する方法を検討したい．

一方，単語の分散表現に，LDA を適用した際のトピック

ⓒ 2019 Information Processing Society of Japan

の生成確率を用いた場合，精度 53.7%（適合率 15.0%，再現率 56.1%），上位 5 件の MAP が 50.0%となり，word2vec による意味類似性を捉えた分散表現が，CNN の判断に有効な情報となっていると言える．

5. 考察

本稿では，仕様情報と不具合情報の組が製品の設計過誤につながる知識として有益かの判断に，非線形な規則表現が必要であると仮定し CNN を適用した．この仮定の妥当性を確認することのみ着目し，4 章の評価とは別に，word2vec による文書の意味類似性と，人間の判断結果について実験を行った．その結果を考察する．

対比手法は，文書中の単語の分散表現を単純加算後，ノルム正規化したものを文書の分散表現とし，その cos 類似度を関連の強さとした．4 章で使用した仕様情報と不具合情報の組をこれに適用した．

その結果，類似度上位 5 件において MAP 60.7%，適合率 33.0%となり，提案手法の精度が高い結果となった．提案手法と対比手法との文書ごとの精度比較を図 **2** に示す．提案手法は，Softmax 層の出力値の上位 5 件中の正解数となっている．文書 4，6, 13，14，17 では対比手法の方が正解数が多い結果となった．対比手法の正解数が多くなる仕様情報の特徴として，単一のトピックのみを記述していることや，他に類似する特徴のないある機器のみについて記述している，という傾向があった．一方，4.2 節で述べたように Softmax 層の出力値 0.9 以上をユーザに提示する場合においても，文書 17 以外に変動はなく，対比手法が得意とする文書の傾向は類似している．つまり，規則の単純性（線形性）が高い文書では，対比手法が有効となっている．これから，2 章で述べたように，本稿で対象とする異なるドメインの内容が混在している文書という前提において仮定は一定の妥当性を有すると考える．文書 12，20 は，いずれの手法においても有効な引き当てができていない．12 は処理フローに関する記載で特徴的な単語が少ない．20 は規格認証の状況を述べた文書で，記載内容が，ハードウェア/ソフトウェア双方のトピックが混在している文書であった．被験者による判定においても有益と判断された組は存在しなかった．このような検索クエリに不向きな文書をどのように検知するかも課題の一つであると考える．提案手法では，文書 12，20 について提示する不具合数が，それぞれ 3 件，1 件であり，Softmax 層出力値による閾値処理により，価値のない結果を効果的に除去できているといえる．

6. まとめと今後の課題

本稿では，仕様情報と不具合情報という異なる性質を持つテキスト情報の組を，宇宙機開発の設計過誤につながる知識として有益かどうか，という観点で検索する手法を提案した．人手による判定が，単純な文章の類似性ではないと仮定し，CNN により判断規則を推定させた結果，実際の宇宙機開発で使用した仕様情報と，不具合情報を用いた検索ユースケースにおいて上位 5 件の MAP 66.4%，適合率 50.0%を達成した．5 章で述べた対比手法のような意味類似性による手法より精度は高いものの，実業務への適用に十分な精度に至ることが難しいことがわかった．また，提案手法は，2 章で述べた 2 つの事象についても改善できた．事象 1 について，類似度上位 5 件の内，仕様文書と無関係な不具合が引き当たる割合が，キーワードマッチングでは 22.0%であったが，提案手法では 6.5%まで改善できた．事象 2 についても、具体例で述べた STT が，スタートラッカと記載された不具合を引き当てることができた．

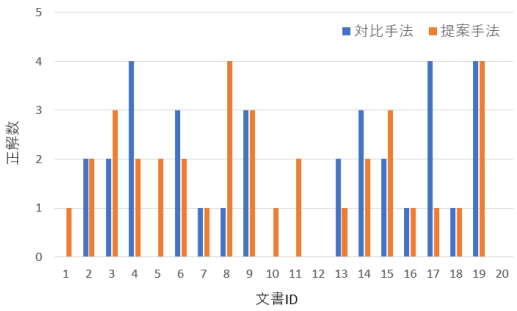

図 2 提案手法と対比手法の精度比較

評価実験では，正例と負例を同数として学習を行ったため，学習データの数は平均して 176 件となった．これは機械学習に用いるデータセットとしては著しく小さいと言える．今後，学習データを増やすことで，提案手法の汎用性向上や，高度なモデルの適用が可能となると考えている．

また，仕様情報として，ある地球周回衛星の特定設計技術領域の範囲としたが，宇宙機の開発では，多様な技術領域の観点からの設計が必要となる．そのため，今回の対象以外の文書への適用は，今後の運用を見据えた上で重要な点であると考えている．データの追加に加え，転移学習なども視野に入れて検討をしたい．

参考文献

[1] Stephen, R. and Hugo, Z,:*The Probabilistic Relevance Framework: BM25 and Beyond*, Journal Foundations and Trends in Information Retrieval, pp. 333 389, 2009.

[2] Yoon, K.:*Convolutional Neural Networks for Sentence Classification*, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[3] 工藤拓，山本薫，松本裕治：Conditional Random Fields を用いた日本語形態素解析，情報処理学会研究報告. 自然言語処理研究会報告，Vol.2004, No.47, pp.8996 (2004).

[4] Tomas, M., Kai, C., Greg, C. and Jeffrey, D.:*Efficient Estimation of Word Representations in Vector Space*, Proceedings of the International Conference on Learning Representations (ICLR 2013).

[5] D, Blei., A, Ng., and M, Jordan,.:*Latent Dirichlet allocation*, NIPS, pp. 601-608, 2001.

ⓒ 2019 Information Processing Society of Japan