7B-01

# 深層学習を用いた摘要情報と伝票項目の対応付けによる伝票の検索精度の向上

藤原新† 石野明†

†株式会社ワークスアプリケーションズ

## 1 はじめに

多くの企業ではその経営資源を管理するために、基幹系システムを導入しており、資源を効率良く活用している。代表的な基幹系システムに会計システムがある。会計システムは企業における会計情報を管理し、関係者に公開する財務諸表を作成するためのシステムである。

企業の会計情報は伝票という形式で会計システムに記録される。一般的に伝票は入力する項目が多く、また、同じ内容の伝票が繰り返し作られる性質がある。そのため、入力担当者は過去の伝票を検索し、そのコピーを元に新しい伝票を作成する場合が多い。

入力担当者は摘要という短いテキストとともに伝票を記録することで、後々伝票を検索しやすくしている。しかし、場合によっては摘要は記入されなかったり、簡略化した表現が用いられるため、文書の類似度に基づく検索技術では目的の伝票が見つからない問題があった。

## 2 関連研究

本稿では伝票の摘要と項目情報を対応付けることで検索の精度を改善する。このように検索語と検索対象を対応付ける検索技術は多く提案されている。Gaoらは LDA[1] を拡張した Bi-Lingual Topic Model により検索語と検索対象の文書を潜在的意味で対応付け、Web 検索の精度を向上させた [2]。また、Huangらは構造的なディープニューラルネットワーク (DNN) による対応付け手法 Deep Structured Semantic Model(DSSM) により、検索精度の向上と計算時間の削減をした [3]。

本稿では Huangらの提案した DSSM を改良し伝票に適用する。Huangらは文書の検索を対象としており、検索語は英単語列、検索対象は英語の文書を使用している。一方、本研究は非言語データである伝票が検索の対象であり、検索語も日本語である。そこで伝票データに合った前処理を加えて行った。

**Improving search quality in Accounting system by using the relation between the title and contents with Deep Learning**

Shin FUJIWARA†, Akira ISHINO†

†HUE&ATE Div. AI Dept., Works Applications Co.,Ltd.

107-6019, Tokyo, Japan


図 1: 一般的な伝票の形式

## 3 伝票形式

伝票は複数の項目からなる非言語かつ、非構造化データである。一般的な伝票を図 1 に示す。伝票はヘッダー部とボディ部からなり、ヘッダー部には摘要、起票日、起票者が記録されている。ボディ部は 2 列に分かれており会計上、左側を借方、右側を貸方と呼ぶ。借方、貸方の両方に 1 つ以上の勘定科目、部署名、金額の項目があり、伝票 1 件で企業内あるいは企業間での 1 回の取引内容が記録される。本研究ではヘッダー部の摘要、ボディ部の勘定科目、部署名の項目群を用いる。

## 4 伝票の項目を考慮した伝票検索

### 4.1 伝票検索のための DNN

伝票の項目群と摘要を対応付けるために、図 2 に示す構造的な DNN を用いる。この DNN は伝票の摘要 $Q$、伝票の項目群 $D_i$ $(1 \leq i \leq N)$ を学習時の入力にとる。$D_i$ には摘要 $Q$ と対応関係のある伝票の項目群 $D^+$ を 1 件と無作為に選択された伝票の項目群 $N-1$ 件を含んでいる。$x_D$、$x_Q$ は項目群と摘要の特徴ベクトル、$y_D$、$y_Q$ は DNN が出力した分散表現のベクトルである。摘要 $Q$ と項目群 $D$ の関連度 $R(Q,D)$ は、分散表現ベクトルのコサイン類似度から計算する (式 1)。

$$R(Q,D) = cosine(y_Q, y_D) = \frac{{y_Q}^T y_D}{\|y_Q\| \|y_D\|} \quad (1)$$

次に摘要 $Q$ と項目群 $D$ が関連する確率 $P(D,Q)$ を

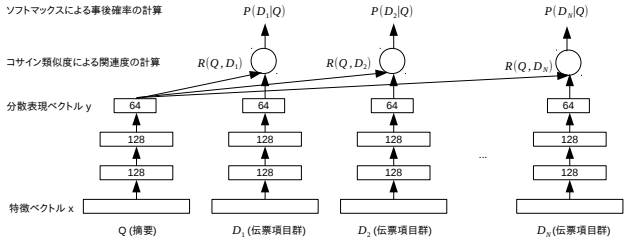


図 2: 伝票検索のための構造的 DNN

ソフトマックス関数で計算する (式 2)。

$$P(D|Q) = \frac{exp(R(Q, D))}{\sum_{D' \in \mathbf{D}} R(Q, D')} \quad (2)$$

対応する摘要と項目群の確率を最大にするため、式 3 の誤差関数が最小になるように学習させる。

$$L(\Lambda) = -log \prod_{(Q,D)} P(D^+|Q) \quad (3)$$

なお本稿では Keras[4] を用いて DNN を実装した。

### 4.2 ベクトル化

伝票の項目は借方、貸方どちらに出現するかで会計上の意味が異なる。そこで項目群の特徴ベクトルは借方貸方の項目を分け、項目を含んでいるか否かで 0,1 のベクトルを作成し、連結して作成した。

伝票の摘要は日本語で特に会計情報に偏った語彙が使われている。DSSM で用いている WordHash は英語に対して高い性能で次元圧縮ができるが、文字の種類が多い日本語では効果的でないことが事前検証で分かった。そこで、語彙外の単語を含む日本語に適した分かち書き手法 Wordpiece[5] により単語を分割し、ベクトル化した。

## 5 実験

### 5.1 実験設計

提案手法が従来の検索技術と比べて有用であるか検証実験を行った。対象のデータは企業 A における 2009 年度から 2010 年度の伝票 32,886 件である。2009 年度の伝票 16,632 件を学習データ、2010 年度の伝票 16,254 件をテストデータとした。実験は “入力担当者が前年度の伝票を検索して新しい伝票の参考にする” ことを想定し、次のように設計した。1. テストデータの伝票から摘要を抜き出し、検索語とする。2. 検索語を用いて学習データから伝票を検索する。3. 元の伝票と結果の伝票の項目群が一致しているか評価する。従来手法として LDA[1] と LSI[6] による検索手法も同様に検証し、精度を比較した。

表 1: 検索精度の比較結果

| | Top1 | Top5 | Top10 | MRR |
|---|---|---|---|---|
| LDA | 18.9% | 34.7% | 42.5% | 0.25 |
| LSI | 31.7% | 45.5% | 50.6% | 0.36 |
| 提案手法 | 43.5% | 46.2% | 47.8% | 0.45 |

### 5.2 実験結果

検索結果の上位 1, 5, 10 件に目的の伝票が得られた割合と MRR を表 1 に示す。上位 1 件では、LDA や LSI は 18%〜31%であるのに対し、提案手法は 43.3%で提案手法の方が高い精度で目的の伝票を得られた。また、検索全体の精度を表す指標である MRR も提案手法の方が高い結果となった。加えて提案手法の検索結果には、会計用語を考慮した結果が見られた。会計において過去の伝票を訂正、削除する場合には、貸方借方を反転した伝票を作成する。“戻し” や “訂正” などの訂正や削除を意味する検索語が与えられたとき、提案手法では反転した伝票を出力しており、会計知識を一部考慮に入れた検索ができていた。

上位 10 件の場合において、提案手法の精度は LSI よりも低くなった。提案手法では、検索対象に同じ項目群を持つ伝票が複数件ある場合、その伝票を連続して提示する性質がある。そのため、上位に総じて同じ伝票を提示してしまい、精度が低くなった。

## 6 おわりに

本稿では DSSM を応用した伝票の検索手法を提案し、精度の検証を行った。検証の結果、提案手法は従来手法と比べて精度の高く、会計知識に基づいた検索ができた。今後、提案手法と従来の手法と組み合わせることにより精度の高いモデルを作成する予定である。

## 参考文献

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:2003, 2003.

[2] Jianfeng Gao, Kristina Toutanova, and Wen-tau Yih. Clickthrough-based latent semantic models for web search, 2011.

[3] Po-Sen Huang, Xiaodong He, et al. Learning deep structured semantic models for web search using clickthrough data, October 2013.

[4] François Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

[5] Yonghui Wu, Mike Schuster, et al. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016.

[6] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407, 1990.