CBOW言語モデルを用いた契約用語の校正手法

山腰 貴大[1]　小川 泰弘[2,3]　中村 誠[4]　外山 勝彦[2,3]
[1]名古屋大学 大学院情報科学研究科　[2]同 情報基盤センター　[3]同 大学院情報学研究科
[4]同 大学院法学研究科
yamakoshi@kl.i.is.nagoya-u.ac.jp

1　はじめに

契約書は，物やサービスの売買，財産の貸借，従業員の雇用などの契約を締結する際に，その内容を表示する書類である．昨今における経済活動の多様化により，契約書を作成する機会は増加の一途をたどっている．しかし，契約書を書き上げることは法律家であっても容易なことではない．契約書を書いた際，提示すべき事柄を法律に則って過不足なく記載しているかどうかだけでなく，契約書特有の専門用語（契約用語）を適切に使用しているかどうかを精査しなくてはならない．しかし，契約書を精査することは専門知識や労力の必要な作業である．また，紙媒体で契約書を作成・精査することが日常的に行われている．

そこで，AI を利用したデータ駆動型の契約書作成支援システムの開発を目指し，その一歩として，使い分けのある契約用語を校正する手法を提案する．具体的には，出現した契約用語の前後の文脈からその契約用語の使い方が適切かどうかを判定し，より適切な契約用語がある場合はそれを提示する．本手法は，このタスクを選択肢つき穴埋め問題とみなし，Continuous Bag-of-Words(CBOW) [1] を用いて解決する．その際，契約書の特徴を加味し，より多くの文脈を扱うために，オリジナルの CBOW に三つの工夫を加える．

本手法と類似したアプローチとして，Long Short-term Memory (LSTM) を用いた文書校正サービスがすでに存在する[1]．しかし，LSTM は入力系列を一つずつ順番に処理するため，長い単語列の処理には多大な時間を要する．一方，CBOW は入力系列をまとめて処理するため，LSTM と比べて高速に動作する．

2　問題の概要と定義

契約用語には互いに似た意味を持つものが数多くあるが，使い分けの基準 [2, 3, 4] が存在するため，文脈に応じて正しく使い分ける必要がある．本稿では，これを契約用語の使い分け問題と呼ぶ．

たとえば，三つの契約用語「者」と「物」と「もの」について考える．契約書においては，言及する事物が人や法人の場合は「者」，その他の有体物の場合は「物」，抽象的な概念の場合は「もの」を用いるという使い分けがある．実際の使用例を以下に示す．

[1]https://prtimes.jp/main/html/rd/p/000000005.000029828.html

者 …解除された者は、解除により生じる損害について、その相手方に対し一切の請求を行わない。
物 …本件建物上に乙が設置した物を乙の費用で全て収去し、…
もの …理由のいかんを問わず本件契約は解除されたものとみなす。

このとき，「…理由のいかんを問わず本件契約は解除された物とみなす。」のように別の用語を使用することは適切でなく，修正する必要がある．この例以外にも，「清算」と「精算」，「課程」と「過程」，「規準」と「基準」，「金額」と「料金」，「違約金」と「損害金」，「過誤」と「瑕疵」と「欠陥」と「欠格」といった契約用語の使い分けが存在する．

本手法では，上述した契約用語の使い分け問題を以下の問題として定義する．

- 使い分けのある契約用語の集合（以降，契約用語の集合）からなる集合 C が与えられる．すなわち，$C = \{T \mid T = \text{契約用語の集合}\}$ である．
- 文 $S = w_1 w_2 \dots w_{|S|}$ 中に現れる契約用語 $t \in T \in C$ が適切かどうか判定する．t は複数の単語からなる場合も想定する．
- より適切な契約用語 $t' \in T$ $(t' \neq t)$ がある場合，それを提示する．

この枠組みによって，本手法は，契約用語の使い分け問題を選択肢つき穴埋め問題とみなす．この問題に関連した研究として，田上 [5] らは大学入試の穴埋め問題を解答する手法を開発した．この手法では，問題のカテゴリを同定する目的で CBOW を用いており，CBOW から解答を直接予測しない．一方，本手法は CBOW から最適な契約用語を直接予測する．

3　CBOW

本手法は，CBOW[1] によって構築した言語モデルを用いて契約用語の使い分け問題を解決する．CBOW は，ニューラルネットワークの一種であり，図 1 に示すネットワーク構造を持つ．入力層は，注目語 w_i の前後 D 単語からなる単語列 $W = w_{i-D} \dots w_{i-2} w_{i-1} w_{i+1} w_{i+2} \dots w_{i+D}$ を入力とし，単語の意味を含んだベクトルに変換する．なお，入力単語数（この場合 $2D$）をウィンドウ幅という．投影層は，全入力単語の単語ベクトルの平均をとる．すなわち，単語 w_k に対応する単語ベクトルを $\boldsymbol{v}(w_k)$ とする

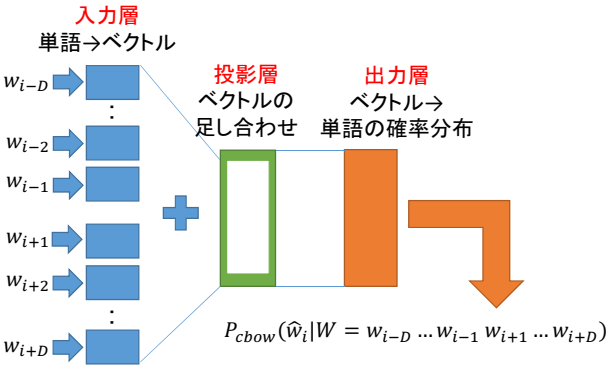

図 1: CBOW

と，W を入力したときの投影層の値 $\boldsymbol{p}$ は以下のとおりである．

$$\boldsymbol{p}(W) = \frac{\sum_{w_k \in W} \boldsymbol{v}(w_k)}{2D} \tag{1}$$

出力層は，投影層の値から注目語 w_i を予測し，w_i に対する予測語 $\hat{w}_i$ の確率分布 $P_{cbow}(\hat{w}_i|W)$ を出力する．

本手法において CBOW を採用する理由は，意味を含んだ単語ベクトルによって文脈情報を扱えるためである．また，LSTM などの RNN による言語モデルと比べて高速に動作すると期待できるためである．

4 提案手法

本手法は次の手順により，契約用語の使い分け問題を解決する．

1. 入力文からの契約用語の抽出
2. 契約用語を別の候補に置き換えた文の生成
3. CBOW 言語モデルによる文の流暢性の計算
4. より適切な契約用語の提示

4.1 入力文からの契約用語の抽出

入力文 $S^0 = w_1^0 w_2^0 \ldots w_{|S|}^0$ を走査し，契約用語の集合 $T \in C$ に属する契約用語 $t^0 = w_i^0 w_{i+1}^0 \ldots w_j^0 \in T$ と S^0 における t^0 の位置 (i_0, j_0) をすべて抽出する．

S^0 の例として，以下の単語列 S_E^0 を考える．

$$S_E^0 = \overset{1}{\text{前項}}\ \overset{2}{\text{に}}\ \overset{3}{\text{定める}}\ \overset{4}{\text{登記}}\ \overset{5}{\text{手続}}\ \overset{6}{\text{に}}\ \overset{7}{\text{要する}}\ \overset{8}{\text{費用}}\ \overset{9}{\text{は}}\ \overset{10}{\text{義務}}\ \overset{11}{\text{の}}\ \overset{12}{\text{履行}}\ \overset{13}{\text{を}}\ \overset{14}{\text{怠っ}}\ \overset{15}{\text{た}}\ \overset{16}{\underline{\text{もの}}}\ \overset{17}{\text{の}}\ \overset{18}{\text{負担}}\ \overset{19}{\text{と}}\ \overset{20}{\text{する}}\ \overset{21}{\text{。}} \tag{2}$$

契約用語の集合を $T_E = \{\text{もの}, \text{者}, \text{物}\}$ とするとき，S_E^0 には契約用語「もの」$\in T_E$ が出現するため，「もの」とその位置 $(16, 16)$ を抽出する．

4.2 契約用語を別の候補に置き換えた文の生成

S^0 中の (i_0, j_0) に出現する t^0 を別の契約用語候補 $t' \in T$ に置き換えた文 S' を生成する．すなわち，

$$S' = replace(S^0, (i_0, j_0), t') \tag{3}$$

である．ここで，関数 $replace$ は以下の式で定義する．

$$\begin{aligned} &replace(S, (i, j), t) \\ &= w_1 w_2 \ldots w_{i-1}\, t\, w_{j+1} w_{j+2} \ldots w_{|S|} \end{aligned} \tag{4}$$

前述の文 S_E^0 の場合，T_E の各契約用語のうち t_E^0 以外のもので t_E^0 を置き換えた次の二つの文を新たに生成する．

$$\begin{aligned} S_E^1 =& replace(S_E^0, (16, 16), \text{者}) \\ =& \text{前項に定める登記手続に要する費用は} \\ & \text{義務の履行を怠った}\underline{\text{者}}\text{の負担とする。} \end{aligned} \tag{5}$$

$$\begin{aligned} S_E^2 =& replace(S_E^0, (16, 16), \text{物}) \\ =& \text{前項に定める登記手続に要する費用は} \\ & \text{義務の履行を怠った}\underline{\text{物}}\text{の負担とする。} \end{aligned} \tag{6}$$

4.3 CBOW 言語モデルによる文の流暢性の計算

前節における置換前の文 S^0 と置換後の各文 S^k の流暢性 flu を CBOW 言語モデルにより計算する．本手法では，S 中の位置 i_0 にある単語 w_{i_0}，つまり，置換前や置換後の文に現れる契約用語 t の先頭の語を注目語とし，CBOW 言語モデルが出力した w_{i_0} の確率を流暢性とする．すなわち，flu は以下の式で求める．

$$flu(S, i_0) = P_{cbow}(w_{i_0}|W) \tag{7}$$

ここで，W は CBOW に入力する単語列である．CBOW のウィンドウ幅を $2D$ とするとき，W は以下のとおりである．

$$W = w_{i_0-D} \ldots w_{i_0-2} w_{i_0-1} w_{i_0+1} w_{i_0+2} \ldots w_{i_0+D} \tag{8}$$

なお，w_{i_0} の前後の単語数が D に満たないときは，空白を表すメタ文字でパディングする．

前述の例においてウィンドウ幅を 4 $(D = 2)$ としたとき，各文の流暢性は次のとおりである．

$$flu(S_E^0, 16) = P_{cbow}(\text{もの} \mid \text{怠っ た の 負担}), \tag{9}$$

$$flu(S_E^1, 16) = P_{cbow}(\text{者} \mid \text{怠っ た の 負担}), \tag{10}$$

$$flu(S_E^2, 16) = P_{cbow}(\text{物} \mid \text{怠っ た の 負担}) \tag{11}$$

ここで，契約書の特徴を加味し，より多くの文脈を扱うために，CBOW に対して次の三つの工夫を施す．

(1) 契約書類型の入力への追加
本手法では，個々の契約書は何らかの類型に属し，各類型に属する契約書には共通性が存在すると仮定する．そこで，表 1 に示す 24 個の類型によって契約書を分類し，各契約書が属する類型 c を入力に加える．類型 c と式 (8) の単語列 W をこのモデルに入力したときの投影層の値 $\boldsymbol{p}_{\text{類型}}$ は以下

表 1: 契約書類型

大分類	小分類			大分類	小分類		
1 物	A 譲渡型	B 貸借型	C 請負型	5 サービス	A 請負型	B 業務委託	C 委任型
2 債権	A 発生型	B 担保型	C 消滅型	6 労働	A 労働	B 就業規則等	C 出向・派遣
3 知財	A 譲渡型	B 利用型	C 請負型	7 組織	A 出資型	B 提携型	C 譲渡型
4 情報	A 秘密型	B 創出型	C 利用型	8 紛争	A 変更	B 解除	C 解決

のとおりである．

$$\boldsymbol{p}_{\text{類型}}(W, c) = \frac{\boldsymbol{v}(c) + \sum_{w_k \in W} \boldsymbol{v}(w_k)}{2D + 1} \quad (12)$$

(2) LR ベクトルの導入

オリジナルの CBOW[1] では，単語ベクトルの平均から投影層の値を求めるため，語順を完全に無視している．それに対して，語順の情報を使用するために，有賀ら [6] は，注目語 w_i の前後で個別に平均を取り，その二つのベクトルを結合したものを投影層の値とする LR モデルを提案した．さらに，Kiryu ら [7] は，このモデルを固有表現検出タスクに利用し，その性能を実証している．後述する LR 行列と区別するために，本稿ではこの工夫を LR ベクトルと呼ぶ．式 (8) の単語列 W をこのモデルに入力したときの投影層の値 $\boldsymbol{p}_{\text{ベクトル}}$ は以下のとおりである．

$$\boldsymbol{p}_{\text{ベクトル}}(W) = \frac{\sum_{k=i_0-D}^{i_0-1} \boldsymbol{v}(w_k)}{D}; \frac{\sum_{k=i_0+1}^{i_0+D} \boldsymbol{v}(w_k)}{D} \quad (13)$$

ここで，; はベクトルの結合を表す．

(3) LR 行列の導入

オリジナルの CBOW[1] では，1 種類の埋め込み行列をすべての単語に対して適用する．それに対して，注目語の前後で異なる埋め込み行列を用いる LR 行列を提案する．これによって，前述の LR ベクトルとは異なるアプローチで語順の情報を使用できる．式 (8) の単語列 W をこのモデルに入力したときの投影層の値 $\boldsymbol{p}_{\text{行列}}$ は以下のとおりである．

$$\boldsymbol{p}_{\text{行列}}(W) = \frac{\sum_{k=i_0-D}^{i_0-1} \boldsymbol{v}_l(w_k) + \sum_{k=i_0+1}^{i_0+D} \boldsymbol{v}_r(w_k)}{2D} \quad (14)$$

ここで，$\boldsymbol{v}_l(w_k)$ は単語 w_k が注目語の前方に現れるときの単語ベクトルであり，$\boldsymbol{v}_r(w_k)$ は w_k が注目語の後方に現れるときの単語ベクトルである．

オリジナルの CBOW にこれら三つの工夫を施して改良したモデルを図 2 に示す．

4.4 より適切な契約用語の提示

t^0 を別の契約用語候補 $t' \in T$ に置き換えた単語列の流暢性が最も高い場合，すなわち，以下の式が成り立つ場合，t' に修正すべきである旨の情報を出力する．

$$t \neq \arg\max_{t' \in T} flu(replace(S^0, (i_0, j_0), t'), i) \quad (15)$$

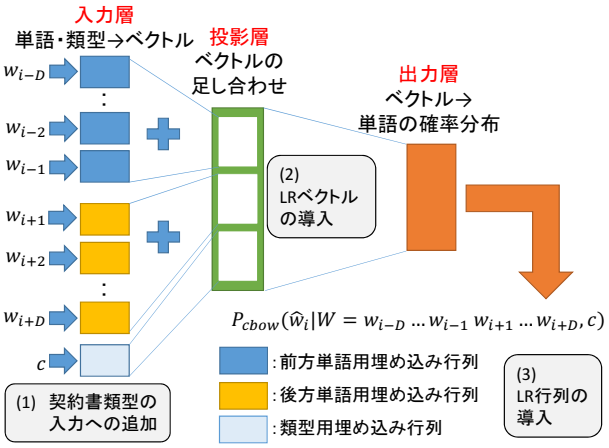

図 2: 改良した CBOW

5 実験

提案手法の性能を評価するために，契約用語の使い分け問題に関する実験を実施した．

5.1 実験設定

実験では 4.3 節で述べた三つの工夫の有無を組み合わせた 8 種類の CBOW 言語モデルを使用した．いずれのモデルも，Chainer(v1.15.0)[2] によって実装し，単語ベクトルの次元数を 200 次元，ウィンドウ幅を 20，エポック数を 10 として学習させた．学習に用いたコーパスは，Web や書籍から取得した 888 件の契約書（57,061 文・1,403,567 単語）から作成した．janome[3] によって契約書を形態素解析し，契約書に出現した 8,710 種類の語すべてを有効な語彙とした．使い分けのある契約用語は文献 [2, 3, 4] から取得し，契約用語の集合の数 $|C|$ は 76，契約用語の総数は 178 となった．

15 件の契約書でオープンテストを実施し，無作為に抽出した 3 件の契約書でクローズドテストを実施した．各手法を用いて，契約書中に出現したすべての契約用語（オープンテストでは 3,216 語，クローズドテストでは 415 語）に対する使い分け問題を解き，その正解率を測定した．

比較のために，LSTM 言語モデルを同様の条件で学習し，CBOW 言語モデルと同様の方法によって性能を評価した．LSTM の隠れ層は，650 次元のものを 2 層設けた．実験における LSTM 言語モデルへの入力方法として，注目語とその前後 10 単語を入力する場合と，単語列全体を入力する場合の 2 通りを試した．

[2] http://chainer.org/
[3] http://mocobeta.github.io/janome/

表 2: 各手法の正解率

手法		Open	Closed
CBOW	オリジナル	83.9%	89.4%
	類型	82.9%	88.4%
	ベクトル	84.9%	90.8%
	行列	84.4%	91.3%
	類型＋ベクトル	84.8%	92.5%
	類型＋行列	85.1%	91.3%
	行列＋ベクトル	85.4%	**93.7%**
	類型＋行列＋ベクトル	**86.5%**	93.3%
LSTM	注目語と前後 10 単語	73.3%	76.1%
	単語列全体	68.2%	72.8%

5.2 結果

実験の結果を表 2 に示す.「類型」,「ベクトル」,「行列」はそれぞれ契約書類型の入力，LR ベクトルの導入，LR 行列の導入を表す．オープンテストではすべての工夫を施したモデルが，また，クローズドテストでは LR ベクトルと LR 行列を導入したモデルがそれぞれ最も高い正解率を示した．また，LSTM 言語モデルを用いたいずれの手法も CBOW 言語モデルを用いた手法の性能に届かなかった．

5.3 考察

すべての工夫を導入したモデルが契約用語の使い分け問題に正答した例を以下に示す．

> の とおり 契約 （ 以下 「 本 契約 」 と ___ 。 ） を 締結 する 。

この例における下線部の候補は「言う」と「いう」であり，CBOW 言語モデルが出力した確率はそれぞれ 5.12×10^{-13}, 4.10×10^{-1} であった．この例の “イウ” は「発言する」という意味を持たないため,「いう」が適切であるが，本手法はこれを正しく判断した．

一方，同じモデルが契約用語の使い分け問題に誤答した例を以下に示す．

> 契約 が 期間 満了 又は 解除 により 終了 した ___ に 存在 する 個別 契約 について は ， 引き続き 本 契約

この例における下線部の候補は「とき」と「時」と「場合」であり，確率はそれぞれ 1.00×10^{-3}, 6.33×10^{-5}, 6.71×10^{-2} であった．正解データに現れる用語は「時」であるが，モデルは「場合」が最も適切だと判断した．その理由として，コーパスにおいて「場合」が「時」よりも頻出するため（それぞれ 8,102 回と 1,123 回）,「場合」がより尤もらしいと判断したことが考えられる．

次に，各モデルにおける処理時間を計測した．1 件の契約書を無作為に抽出し，単語列の流暢性に対する平均計算時間を計測した．処理は GPU (NVIDIA Tesla K40c) 1 枚を用いて行った．この操作を 10 回行ったときの平均値と最大値を表 3 に示す．CBOW 言語モデルと比較すると，LSTM 言語モデルの処理時間は，入力単語列を注目語とその前後 10 単語にした場合に 10 倍以上，単語列全体を入力とした場合に 20 倍以上となった．この結果より，CBOW 言語モデルによる手法は，LSTM 言語モデルによる手法と比べて高速に処理できることが明らかとなった．

表 3: 各モデルの処理時間 [msec]

手法		平均	最大
CBOW	オリジナル	7.20	7.27
	類型	7.56	7.67
	ベクトル	7.69	7.81
	行列	7.35	7.46
	類型＋ベクトル	8.07	8.24
	類型＋行列	7.54	7.65
	行列＋ベクトル	7.62	7.79
	類型＋行列＋ベクトル	8.06	8.20
LSTM	注目語と前後 10 単語	101	101
	単語列全体	176	180

6 まとめ

本論文では，使い分けのある契約用語を校正する手法について述べた．本手法は，三つの工夫を施した CBOW 言語モデルを用いて契約用語の使い方を判定し，より適切な用語がある場合それを提示する．実験の結果，LSTM 言語モデルの 10%以下の時間で処理し，コーパス外の契約書で 86.5%，コーパス内の契約書で 93.3%の正解率を達成することが分かった．

今後は，モデルの性能向上を図ると同時に，判定対象の契約用語を自動抽出する仕組みを導入したい．

謝辞

本研究を実施するにあたり，表 1 の契約書類型をはじめ，契約書に関する様々な知識をご提供いただいた山田尚武代表弁護士ほか弁護士法人しょうぶ法律事務所の皆様に深く感謝する．また，本研究は，名古屋大学博士課程教育リーディングプログラム「実世界データ循環学リーダー人材養成プログラム」の支援を受けて行ったものである．

参考文献

[1] T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representation in Vector Space. In *ICLR*, 12 pages, 2013.

[2] 高橋均，稲田博志．契約用語使い分け辞典．新日本法規出版，2011.

[3] 川崎政司．注釈公用文用字用語辞典〔第七版〕．新日本法規出版，2015.

[4] 礒崎陽輔．分かりやすい公用文の書き方 改訂版．ぎょうせい，2010.

[5] 田上諒，木村輔，宮森恒．大学入試の穴埋め型問題に対する語順を考慮した自動解答手法．情報処理学会論文誌データベース，Vol. 10，No. 3，pp. 45–57，2017.

[6] 有賀竣哉，鶴岡慶雅．単語のベクトル表現による文脈に応じた単語の同義語拡張．言語処理学会第 21 回年次大会発表論文集，pp. 752–755，2015.

[7] Y. Kiryu, A. Ito, K. Yamasawa, T. Kasahara. Effective Neural Network Models for Predicting Unprintable Words in Japanse Judicial Precedents. Proc. of the 11th *JURISIN*, pp. 36–47, 2017.