4G1-2

文の位置関係と意味情報を用いた少量文書からの知識構造化

Structured data extraction from a small number of documents by using semantic and structural features between sentences

本間 幸徳 Yukinori Homma　貞光 九月 Kugatsu Sadamitsu　東中 竜一郎 Ryuichiro Higashinaka　浅野久子 Hisako Asano　松尾義博 Yoshihiro Matsuo

日本電信電話株式会社 NTT メディアインテリジェンス研究所
NTT Media Intelligence Laboratories, NTT Corporation

In this paper, we propose a method for structured data extraction from a small number of documents. The proposed method is composed of a meaning understanding model and a structural understanding model. The proposed method learns to extract sentences about attributes by using semantic and structural features between sentences and generates a structured database from the extracted sentences. Experimental results show the proposed method outperforms baseline methods.

1. はじめに

近年，管理や保存のために大量の文書の電子化が進んでおり，電子化された文書から必要な情報を抽出し活用する技術への期待が高まっている．例えば保険商品に関する複数のパンフレットから保険に関する知識を抽出しデータベースを構築することができれば，ユーザの保険商品に関する質問や商品間の比較・問い合わせなどに的確に回答することが可能になる．

複数の商品に関する文書が与えられた時，データベースを構築するためには，商品の属性に関するテキストを抽出する技術が必要となる．属性とはあるドメインの商品が共通して持つ知識のことで，保険商品であれば「保険対象」や「保険料」等が該当する．

本稿では HTML 構造を持つ保険商品の説明文書から，属性に関する知識を記述しているテキストである「値」と，条件や場合を示すテキストである「条件」を抽出し，抽出したテキストから構造化された知識を獲得することを目的とする．ここでは HTML 構造で区切られる文単位のテキストを抽出の対象とする．例えば，図 1 に「商品 A」に関する説明文書の例を示す．図 1 から属性「保険対象」に関するテキストの抽出を考える時,「条件」として「個人コース」が,「値」として「加入した方のみが保険の対象となります」が，抽出できる．抽出された条件と値の対応関係を取る事で,「商品・属性・条件・値」からなる構造化された知識を獲得することができる．図 1 の文書から構築されるデータベースの例を表 1 に示す．

文書から特定のテキストを抽出する情報抽出の従来手法としては，文書の構造パタンを利用して抽出する手法 [Auer 07, 玉川 11] と文の意味情報を利用して抽出する手法 [Nagy 12, Joshi 15] がある．前者は文書構造から特定のパタンを作成し，パタンに適合するテキストを抽出する手法で，抽出したいテキストが共通のパタンで記述されている場合に効率的にテキストを抽出できる．しかしながら，今回対象とする文書の一部はリスト構造や見出し文等の特徴的な文書構造で記述されているものの，その他多数のテキストは多様な HTML 構造の組み合わせで記述されており，適切なパタンを獲得することは難しい．後者の手法はあらかじめ訓練文書から特定の属性に関して用いられる単語を獲得し，対象とする文書から似た単語を持つテキストを抽出する手法である．多数の学習文書が与えられた時に

連絡先: 本間 幸徳，NTT メディアインテリジェンス研究所，神奈川県横須賀市光の丘 1-1，homma.yukinori@lab.ntt.co.jp

```
<div>
 <h4>■引受条件</h4>
 <ul>
 ・・・

  <li>●被保険者</li>
 会員本人、または配偶者、同居の親族及び別居の未婚の子を被保険者としてご加入いただけます。
 <ul>
  <li>【夫婦コース】</li>
 申込人が加入すれば、申込人の配偶者も保険の対象となります。
  <ul>
   <li>※続柄は事故発生時におけるものをいいます。</li>
  </ul>
  <li>【個人コース】</li>
 加入した方のみが保険の対象となります。

 . . .
  </ul>
</div>
```

図 1: 抽出対象の文書例

表 1: 構築されるデータベース例

商品	属性	条件	値
商品 A	保険対象		会員本人、または配偶者、同居の親族及び別居の未婚の子を被保険者としてご加入いただけます。
商品 A	保険対象	夫婦コース	申込人が加入すれば、申込人の配偶者も保険の対象となります。
商品 A	保険対象	個人コース	加入した方のみが保険の対象となります。

特に有効な手法であるが，訓練文書が比較的少量の場合は学習する意味情報の不足が課題となる．

本稿では，比較的少量の保険商品の説明文書を用いて構造化された知識を獲得する手法を新たに提案する．提案手法は，多様な構造を持つ文書に対して，文書を文の系列と見做し系列ラベリングの手法を用いることで，リスト構造等の特徴的な文書構造を考慮しつつ，特定のパタンに依存せずにテキストを抽出する．また少数の訓練文書から単語情報を効率的に獲得するために，あらかじめ入力文書以外のコーパスを加えて学習した単語の分散表現を意味情報として用いる．単語の分散表現を利用することで，表層が異なるが意味が類似した単語から学習することが可能となり，意味情報の不足を補うことが期待できる．本稿では文単位のテキストを抽出するために，単語の分散表現

を文単位に拡張した手法を用いて学習を行う．抽出された値と条件に関するテキストについて，ルールに基づいて対応関係を取ることで構造化された知識を獲得する．

2. 関連研究

文書から特定の属性に関する情報を抽出の従来手法としては，文書の構造パタンを利用して抽出する手法 [Auer 07, 玉川 11, Muslea 99, Gulhane 11] と文の意味情報を利用して抽出する手法 [Nagy 12, Joshi 15] がある．

人手で作成したパタンを用いて大規模な知識ベースを作成した代表的な例として DBpedia [Auer 07] がある．DBpedia では主に infobox の構造を利用したパタンを用いることで，英語 Wikipedia から大規模なデータベースを構築した．玉川ら [玉川 11] は日本語 Wikipedia から情報を抽出する際に Wikipedia 本文中のリスト構造を考慮したパタンを用いる手法を提案している．人手によるパタンを用いる手法は Wikipedia のように統一された構造を持つ文書が大量に存在する場合は効率的に情報を抽出できる一方，多様な構造を持つ文書に対しては網羅的にパタンを書ききることは難しいという問題がある．パタンを自動的に獲得するために，機械学習を用いる手法も多く提案されている [Muslea 99, Gulhane 11]．Muslea らは半構造化文書のタグ構造を木構造と見做して木構造のノード間のパスをパタンとして学習する手法を提案している．Gulhane らは多様な構造を持つ Web 文書から適切に情報を抽出するために，Web 文書の構造種別をクラスタリングして種別ごとに適切なパタンを割り当てる手法を提案している．自動的にパタンを獲得する手法は，特に通販サイトや製品サイト等のテンプレート的な構造を多数持つ Web ページに対して，情報抽出に適切なパタンの獲得が容易である一方，比較的少量な文書群に対しては，パタンを自動で獲得するための情報が不足するために精度良く知識を抽出することは困難である．

文の意味情報を用いて情報を抽出する手法の研究として Nagy ら [Nagy 12] の研究がある．彼らは特定の人名をクエリとした検索結果の Web ページから職業や電話番号等の人物属性に関するフレーズを，固有表現抽出器によって抽出する手法を提案している．彼らの手法では意味情報として単語の表層情報を主に用いており，訓練文書に出現しない単語について学習することが難しい．近年では単語の意味情報として入力文書外のコーパスから学習した単語の分散表現を用いる手法が提案されており，情報抽出にも用いられている [Joshi 15]．単語の分散表現を利用することで，表層が異なったとしても意味が類似した単語から学習することが可能となり，意味情報の不足を補うことが期待できる．Joshi らは e コマースに関する固有表現抽出に関して，単語の分散表現を用いる場合を含むいくつかの条件に関して比較実験をしており，特に大規模なコーパスとドメインに特化した小規模な文書を分散表現の学習に用いた場合に精度が向上することを報告している．本稿でもこの利点を活かし，Web 上で獲得した大規模なコーパスと保険に関する比較的少量の文書を単語の分散表現の学習に用いることで，抽出精度の向上を図る．

3. 提案手法

提案手法は文の意味情報を獲得する意味理解モデルと文書の構造情報を利用して文を分類する構造理解モデルから構成される．

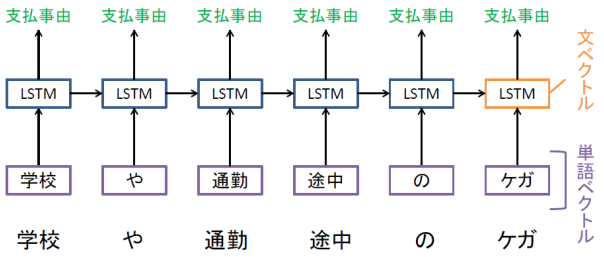

図 2: 意味理解モデルによる文ベクトル生成例

3.1 意味理解モデル

意味理解モデルでは文の意味情報を学習するための素性の獲得を行う．まず文中の各単語について単語の分散表現である単語ベクトルを獲得する．次に獲得された単語ベクトルを用いて文ベクトルに変換する．近年このような文ベクトルへの変換には Recurrent Neural Networks [Mikolov 11] などのニューラルネットワークを利用した手法があり，特に LSTM [Hochreiter 97] はその表現能力と単語間の長期的な依存関係を捉える能力が高いことが知られている．

文ベクトル生成の一例を図 2 に示す．まず入力された文を形態素解析し単語系列を得る．図 2 では入力文「学校や通勤途中のケガ」に対して,「学校，や，通勤，途中，の，ケガ」の 5 つの単語が得られる．次に各単語に対して単語ベクトルを獲得する．ここで単語ベクトルとして word2vec [Mikolov 13] による分散表現を用いる．単語ベクトルの学習には，訓練文書を含む保険に関する比較的小規模の文書と，Wikipedia から獲得した大規模なコーパスを用いる．

各単語ベクトルは入力ベクトルとして LSTM モデルに入力される．t 番目の単語を入力した時，当該文が分類される属性を出力するように学習を行う．出力ベクトル y_t は分類される属性の個数と同じ次元数を持ち，以下の式で表現される．

$$y_t = softmax(W^y h_t + b^y). \quad (1)$$

ここで W^y, b^y は出力層に関する重み行列とバイアスベクトルである．また h_t は t 番目の単語を入力した時の LSTM モデルの隠れ層を示す．入力文のすべての単語ベクトルが入力された時刻 T の隠れベクトル h_T を当該文の意味を表現する文ベクトルとして扱い，次節の構造理解モデルで用いる．

3.2 構造理解モデル

本節では前節で作成した文ベクトルと文書の構造情報を用いて，文を属性に分類する構造理解モデルについて説明する．文を属性に分類するためには，文の意味情報に加えてその文の見出し文や周辺の文などの情報も重要な手掛かりになる．構造理解モデルでは文書を文の系列データと見做して系列ラベリング手法を用いることで，文間の関係性を考慮して文を分類する．系列ラベリング手法として長距離の系列間の関係を学習できる LSTM モデルを用いることで，離れて記述された文間の関係性を学習することができる．

図 3 に提案するモデルの概略図を示す．構造理解モデルではまず文の構造的な情報を構造ベクトルとして獲得する．構造ベクトルは文を直接内包する HTML タグの one-hot 表現であるベクトルと，対象としている文と一つ前に出現した文との HTML 文書中での階層の深さの差を情報として持つベクトルからなる．構造ベクトルと前節で作成した文ベクトルを結合し

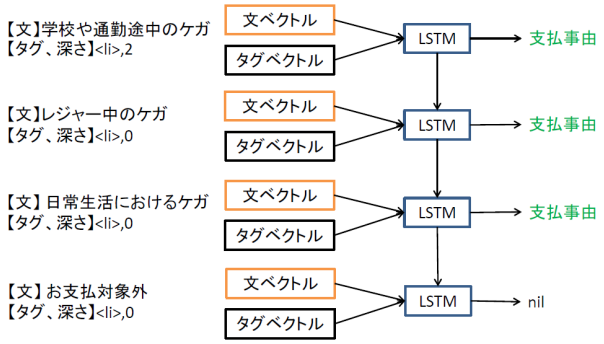

図 3: 構造理解モデルによる文分類の一例

たベクトルを LSTM モデルに入力する．l 番目の文に対して出力ベクトル s_l が当該文が分類される属性を出力するように学習を行う．出力ベクトル s_l は分類されうる属性の個数と同じ次元数を持ち，以下の式で表現される．

$$s_l = softmax(W^s h_l + b^s). \quad (2)$$

ここで W^s, b^s は構造理解部における出力層に関する重み行列とバイアスベクトルである．また h_l は l 番目の単語を入力した時の LSTM モデルの隠れ層を示す．

3.3 知識の構造化

与えられた文書中の全ての文について値・条件に分類された後，以下のルールを満たす場合に，値と条件に分類されたテキスト間に条件関係が有ると対応付ける．

- 対象とするテキストは，HTML 構造における同一の特定のタグ (div) 内に記述されている
- 文書中で，条件に分類されたテキストと値に分類されたテキストの順で記述されており，間に他の条件に分類されたテキストを含まない

条件関係が有ると対応付けられた複数のテキストを，一つの構造化された知識として獲得する．

4. 評価実験

4.1 実験に使用したデータセット

提案手法及び比較手法の学習・評価用データセットとして，7 種類の HTML で記述された保険商品のパンフレットに対して，文書中の各文に人手で属性に関する知識を表す「値」であるか，または「条件」であるか，いずれの属性にも分類されない「属性なし」であるかを示すラベルをアノテートした．ここで一つの文に複数のラベルが付与されることを許した．例えば,「がん先進医療特約」という文は属性「特約」に関する値を示すテキストであると同時に，他の属性の値に関する条件を示すテキストであるとも考えられる．本稿では，複数ラベルのアノテートを許し，学習時は出現回数の多いラベルを選択して学習を行った．

属性の種類は 7 種類の保険商品 (医療保険 4 種類，傷害保険 3 種類) に関する文書から，共通に持たれやすい知識であるかどうかを判断基準として作業者 2 名によって決定した．表 2 に「条件」と「属性なし」を除いた，37 種類の属性一覧を示す．

表 2: 属性の一覧

支払事由	免責事由	保険金請求時必要書類
保険期間	契約解除の扱い	保険料の支払・払込方法
引受限度	保険料払込期間	保険料決定の仕組み
窓口・相談先	通知・告知事項	保険金額・給付金
解約払戻金	支払対象外期間	保険料の払込猶予期間
年齢制限	更新手続き方法	保険会社破綻時の取扱い
申し込み方法	保険の対象	クーリングオフの扱い
特約	割引率	死亡保険金受取人
特徴	商品の正式名称	ペットネーム
解約手続き	引受保険会社	個人情報の取扱
申込締切日	払込免除の有無	満期返戻金
補償開始	補償終了	基本契約
保険料		

4.2 実験方法

本実験ではデータセットに対してそれぞれ学習・評価用データとして用いて，比較するそれぞれの手法がどの程度正しく各データセットを再現できるか，文書中の各文に対する推定ラベルと正解ラベルに関する適合率・再現率・F 値を算出することで評価した．ここで正解ラベルとは各文に人手で付与された 1 つ以上のラベルを示し，推定ラベルとは各手法の出力において最大の確率を示した属性を示す．推定ラベルが正解ラベルに含まれている場合に正解とカウントした．実験は 1 文書を評価用の文書として，残りの文書を学習用の文書として用いる交差検定法によって評価した．

本実験では以下の 4 つの手法に対して比較を行った．

- (a) 単語の表層を特徴量とする最大エントロピーモデル
- (b) 単語の表層を特徴量とする条件付き確率場
- (c) 提案手法 w/o 構造情報
- (d) 提案手法

(a) は文中の単語の表層を特徴量とする最大エントロピーモデル (Maximum entropy model) である．学習用の文書が持つ各文について，形態素解析によって分割した単語の表層を特徴量とした入力ベクトルを用いて各文が分類される属性を出力するように学習を行った．実験では，意味情報のみを考慮するベースライン手法として扱う．(b) は文中の単語の表層を特徴量として，提案手法と同様に文書の出現順に整列した文の系列を学習した条件付き確率場 (Conditional random field) を示す．系列の各要素は形態素解析によって得られた各単語のユニグラム情報を持つ．提案手法と同様に各時刻に対応する文が分類される属性を出力するように学習を行った．実験では，文書の構造情報を考慮するベースライン手法として扱う．(c) は (a) と同様の手法だが，特徴量として文中の単語の表層情報ではなく提案した文ベクトルを入力とした最大エントロピーモデルである．提案手法において文間の関係性を考慮しないモデルに対応する．(d) は 3 章で説明した提案手法を示す．

4.3 実験結果

実験結果を表 3 に示す．提案手法は単語の表層を用いる手法 (b) に比べ F 値で 5 ポイント以上上回っており，単語の意味的な情報を獲得し文ベクトルに変換する手法が効果的であることを示している．出力例を見ると提案手法は，特にフレーズで記述された属性の値に分類されるテキストについて，高精度で分類していた．例えば,「満 30 歳から満 85 歳」というテ

表 3: 実験結果

手法	適合率	再現率	F 値
(a) 最大エントロピーモデル	0.484	0.454	0.462
(b) 条件付き確率場	0.552	0.436	0.469
(c) 提案手法 w/o 構造情報	0.539	0.458	0.484
(d) 提案手法	**0.608**	**0.480**	**0.525**

表 4: 構造化知識の獲得例

商品	属性	条件	値
がん保険	支払事由		がんを治すための「三大治療 (手術・放射線治療・抗がん剤治療)」をしっかり補償
がん保険	支払事由	がん先進医療特約	がんの診断や治療を目的として、所定の先進医療を受けたとき
がん保険	引受限度	①三大治療のための通院	支払日数は無制限

キストは属性「年齢制限」の値に分類すべきだが,「満 · · · 歳」というフレーズは保険文書では多数記述されているため学習時に負例の数が大きく，単語単位で分類する手法では正しい分類先を学習することが難しい．一方で，提案手法は「満 · · · 歳から満 · · · 歳」というフレーズを意味情報として特徴抽出することができたため，正しく分類することができたと考えられる．

また文のみを特徴量として用いる手法 (a)(c) に比べ周囲の文を特徴量として考慮する手法 (b)(d) が特に適合率において高い値を示しており，構造情報を用いることで知識の抽出精度が向上することがわかる．提案手法とベースライン手法 (c) の推定結果を比較すると，特に同じ属性の値がリスト構造として記述されている場合に精度よく抽出できていた．また，特定の属性が記述されることを示すテキストである「見出し」が記述されている場合においても高精度で抽出できていた．保険文書では，見出しと属性の値に関するテキストが離れて記述されることが多く，離れて記述された文間の関係性を学習できる LSTM モデルが有効に機能したと考えられる．

最後に構造化された知識として獲得できた例を表 4 に示す．例えば属性「引受限度」について,「①三大治療のための通院」の場合「支払日数は無制限」であるという知識が獲得できている．文全体を抽出対象としているために,「条件」の出力に「①」等の不要な単語が含まれている．整理された知識獲得のためにはこれらの単語を除く必要があり，形態素ごとの情報抽出が今後の課題となる．

5. おわりに

本稿では比較的少量かつ多様な文書構造を持つ文書群に対して，構造化された知識を獲得するための手法を提案した．大規模なコーパスで学習した文の分散表現を利用することと，文書を文の系列と見做しリスト構造等の特徴的な文書構造を考慮して学習することで，ベースライン手法に比べ高い精度でテキストを抽出できることを示した．

本稿では条件と値の対応付けをルールに基づいて行ったが，関係抽出の手法を用いることで，より高精度に対応関係を獲得することが考えられる．今後の研究課題として取り組んでいきたいと考えている．

参考文献

[Auer 07] Auer, Sören and Bizer, Christian and Kobilarov, Georgi and Lehmann, Jens and Cyganiak, Richard and Ives, Zachary, Dbpedia: A nucleus for a web of open data, Springer (2007)

[玉川 11] 玉川奨，森田武史，山口高平，日本語 Wikipedia からプロパティを備えたオントロジーの構築，人工知能学会論文誌，Vol. 26，No. 4，pp. 504-517 (2011)

[Muslea 99] Muslea, Ion and Minton, Steve and Knoblock, Craig, A hierarchical approach to wrapper induction, Proceedings of the third annual conference on Autonomous Agents, pp. 190-197 (1999)

[Gulhane 11] Gulhane, Pankaj and Madaan, Amit and Mehta, Rupesh and Ramamirtham, Jeyashankher and Rastogi, Rajeev and Satpal, Sandeep and Sengamedu, Srinivasan H and Tengli, Ashwin and Tiwari, Charu Web-scale information extraction with vertex, Data Engineering (ICDE), 2011 IEEE 27th International Conference on, pp. 1209-1220 (2011)

[Mikolov 11] Mikolov, Tomáš and Kombrink, Stefan and Burget, Lukáš and Černockỳ, Jan Honza and Khudanpur, Sanjeev, Extensions of recurrent neural network language model, Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp. 5528-5531 (2011)

[Hochreiter 97] Hochreiter, Sepp and Schmidhuber, Jürgen, Long short-term memory, Neural computation, Vol. 9, No.8, pp. 1735-1780 (1997)

[Mikolov 13] Mikolov, Tomas and Chen, Kai and Corrado, Greg and Dean, Jeffrey, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013)

[Nagy 12] Nagy, Istvn, and Richrd Farkas, Person Attribute Extraction from the Textual Parts of Web Pages, Acta Cybern, Vol. 20, No.3, pp. 419-440 (2012)

[Joshi 15] Joshi, Mahesh and Hart, Ethan and Vogel, Mirko and Ruvini, Jean-David, Distributed Word Representations Improve NER for e-Commerce, Proceedings of NAACL-HLT, pp. 160-167 (2015)