オントロジー形式アノテーションを対象とした交通用語・関係抽出と正誤問題の回答

鈴木 直樹　Bou Savong　三輪 誠　佐々木 裕
豊田工業大学
{sd16042, savong, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

1　はじめに

現在，人類が所有する知識の多くが書籍や論文等の文書情報として存在している．これらに含まれる知識を正確かつ効率的に見つけ出すことができれば，問題解決に活用することができる．しかし，文書情報は多様な表現により曖昧性を含んで記述されている非構造化情報であるため，問題解決システムにおいては，(1) 文書情報中の知識を構造化する機能と (2) 解くべき課題と関連する内容と照合することで解を導く機能が必要である．

交通関係の文書に対する (1) の実現に関しては，文書に用語抽出と関係抽出を適用し，必要な内容を抜き出す方法が研究されてきた．長野 [7] は，BiLSTM と CRF を用いて交通に関する教則文から用語抽出を行うモデルを作成した．八木ら [8] は，Convolutional Neural Network（CNN）を用いて交通に関する教則文に用語タグを付けた文書から，関係抽出を行うモデルを作成した．このようなモデルを利用し，学習用データを用意すれば，文を理解するのに必要な情報を抜き出し，知識のデータベースを作成できるようになることが好ましい．しかし，現状では，これらのモデルの個別の精度は F 値で，用語抽出は 0.593，関係抽出は 0.407 にとどまっている．さらに，知識をデータとして抜き出すには，用語抽出と関係抽出をつないで実行する必要があり，予備実験では 25% 程度しか正しく用語と関係の両方の抽出を行うことができず，データベースの作成には程遠い結果となった．このような結果となる原因として，用語の分類において，似た意味を持つタグが含まれていたこと，関係のアノテーションの分類が細かく関係抽出器における分類を困難にしていたこと，が考えられる．この問題を解決するために，似た意味を持つタグを整理・統一し，関係を用語としてタグ付けした新しいオントロジー形式のアノテーションが有用と考えた．

本研究では，オントロジーのデータ形式を持つ新しいアノテーションデータを用いて，用語抽出・関係抽出の精度を向上し，より高い精度での知識の構造化を目指す．さらに，(2) の実現に向けて，交通に関する問題文に対して，アノテーションしたデータを利用した構造化した知識を用いて解答を得るための，BERT (Bidirectional Encoder Representations from Transformers) モデル [2] を基盤とした正誤問題の回答システムを構築する．

2　関連研究

2.1　Flair

Flair は Alan ら [1] によって開発された自然言語処理用のライブラリである．文を単語ではなく文字の列として扱い，文内の周囲の情報から埋め込みを行うことで学習を行う．これにより，異なる文脈における同じ語彙単語文字列に対して，異なる埋め込みを行うことができる．

2.2　八木らの提案した関係抽出モデル

深層学習により，1 つの用語に関係の始点としての用語表現と終点としての用語表現を作成し，それを利用して用語 1 から用語 2 に向かう関係と用語 2 から用語 1 に向かう関係のスコアを計算する．そしてスコアが高い方をその用語間の関係と予測するモデルである．

2.3　BERT

BERT (Bidirectional Encoder Representations from Transformers) は，Devlin ら [2] によって提案された様々な自然言語処理に関するタスクに適応できるモデルである．ラベルなしデータから作成した事前学習モデルを，目的のタスクに適応するようモデルの微調整を行うことにより，少ない学習データから良い結果を得られるようにしている．BERT は Transformer[3] モデルを Base では 12 回，Large では 24 回重ねた構造を

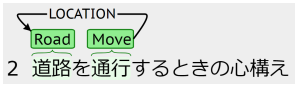

図1　旧アノテーション

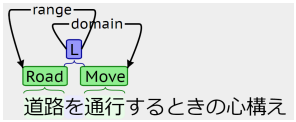

図2　新アノテーション

している．文の比較を行うタスクでは，入力した 2 つの文それぞれの末端に [SEP] のトークンを追加し，入力の先頭に [CLS] のトークンを追加して，全体を 1 つの入力として学習を行っている．事前学習の際に，[CLS] のベクトルを用いて 2 つの文の内容が続いているのかそうでないのかを判定するタスクを解くことで，[CLS] が 2 つの文全体の内容を示すように学習を行っている．

3　使用したアノテーションデータ

本研究で使用した 2 種類の形式のアノテーションについて説明する．従来のアノテーション [4] を図 1に，オントロジー形式の新しいアノテーション[*1]を図 2に示す．

従来のアノテーションでは，文章に交通用語を表す用語タグと，交通用語間の関係を表す関係リンクを使用していた．用語タグの種類には人を表す「Person」や車を表す「Car」などがある．関係リンクは，用語タグで抽出した単語同士がどのような関係を示すのかを表すものであり，用語間の位置関係を示す「LOCATION」や，主体を示す「domain」などがある．

新アノテーションは従来のアノテーションをベースにして作成しており，タグの表現方法を従来のアノテーションから変更したものである．従来のアノテーションから新アノテーションへの変更では，その関係を与える根拠となった文章中の単語に関係名と同じ名前の用語タグ（関係タグ）を与え，その関係タグと従来のアノテーションでは関係リンクと結ばれていた用語を domain, range で結んでいる．従来のアノテーションでは用語タグが 565 種類あったのに対し，似た意味のラベルを統一し，新アノテーションでは 551 種類の用語・関係タグでタグ付けを行った．また，従来のアノテーションではリンクの種類が 47 種類あったのに対し，新アノテーションではリンクの種類を主に domain，range，subClassOf，TIME，TOOL の 5 種類に抑えている．

[*1] 本アノテーションの詳細は言語処理学会第 26 回年次大会の別の発表にて発表予定．

4　提案手法

4.1　用語抽出・関係抽出

オントロジー形式のアノテーションを用いて，用語抽出・関係抽出を行う．従来のアノテーションでは，アノテーションのタグ，特に関係タグの種類が多く，さらには似た意味の関係タグや文書中に一桁の回数しか登場しない関係タグがあり，関係抽出の抽出精度が用語抽出の抽出精度と比べて特に低くなっていた．そこで，関係タグの種類を 5 種類に絞り，関係を用語タグとしてタグ付けし，用語タグの種類も意味が似たものを省いた 551 種類とした新アノテーションを用いて用語抽出・関係抽出を行う．ここで，用語抽出に関しては，事前実験において，長野らが提案したモデルよりも Alan らが提案した Flair の方が高精度に抽出できたため，用語抽出は Flair を用いる．

4.2　正誤問題の回答

オントロジー形式のアノテーションを用いた，交通法に関する正誤問題に回答するモデルを提案する．BERT Base モデルを利用して，問題文と学習文の内容が意味的に同じであるかを判断するタスクを行い，正誤問題の判定を行う．問題文と学習文の比較方法は，文から取り出した交通用語単位での比較である．本研究で使用したモデルは図 3のようになっている．まず，比較したい 2 つの文の間と文末に [SEP] のトークンを追加し，2 つの文の前に [CLS] のトークンを追加する．作成したものを入力として BERT に入力し，入力をベクトル化する．それぞれの文について交通用語のトークンのベクトルを取り出しそのベクトルの平均を文ベクトルとする．このように計算した，問題文と学習文それぞれの文ベクトルの内積をとり，シグモイド関数を掛けて予測を行う．

5　実験

交通法規に関する文書 [5] を対象とした従来の形式のアノテーションコーパスと新しい形式のアノテーションコーパスについて，用語抽出・関係抽出の精度の違いを調べた．また，文献 [6] によって作成された交通法規に関する問題文の内の 78 問を，交通法規に関する文章のデータと比較して正誤判断を行った．事前学習は同文書を使用し，京都大学の黒橋・河原研究室で公開されている「BERT 日本語 Pretrained モデル」を参考にし

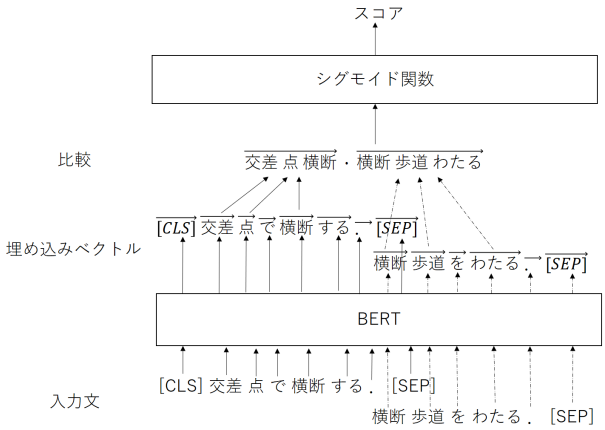

図3　提案モデル

て行った．交通法規に関するコーパスと交通法規に関する問題文のデータの詳細はそれぞれ表 1, 2のようになっている．

5.1　用語抽出・関係抽出の実験

Flair を用いて用語抽出を行った結果をそれぞれ表 3, 表 4にまとめた．表には用語毎の結果の内，文書内での出現回数が多かった 5 つの用語の結果と，その 5 つ以外の抽出結果のマイクロ F 値と，すべての用語抽出結果のマイクロ F 値を示した．実験は 5 分割交差検証によって行った．新アノテーションでは，従来のアノテーションで用語抽出を行った場合と比べて，精度が低くなった．これは，新アノテーションで，従来のアノテーションでは関係として扱っていた内容を用語として扱うようになったことが原因として挙げられる．関係を示す用語も用語抽出で予測する必要が出てきたため，予測する用語が増加・複雑化し，結果的に精度が低くなったと考えられる．用語毎の F 値を確認すると，従来のアノテーションは出現頻度の高い用語の F 値が全用語の F 値を上回っていた．それに対し，新アノテーションは一番出現回数の多い Property の F 値が全体の F 値を下回っていた．新アノテーションの結果でも，Road や Pass の F 値は従来のアノテーションと同様に高い F 値を示しており，新しく追加した関係を表す用語の抽出精度が低いことが分かった．これは，関係を表す用語の定義が複雑で，抽出する難易度が高かったからであると考える．例えば，最も出現回数が多いにもかかわらず F 値が低い Property は，文内で抽出している語が「の」や「が」などの 1 文字であることも多く，予測が困難であると考える．

八木らの提案したモデルを用いて関係抽出を行った結果をそれぞれ表 5，6に示す．表には関係毎の結果の内，文書内での出現回数が多かった 3 つの関係の結果と，すべての関係抽出結果のマイクロ F 値を示した．実

表1　交通法規に関する文章のデータ

アノテーション	従来	新
文数	1,288	1,295
用語数	8,832	12,332
用語ペア数	34,207	76,523
関係のあるペア数	7,071	10,627
用語タグ種類	565	551
関係タグの種類	47	5

表2　問題文のデータ

アノテーション	従来	新
問題数	78	78
用語数	460	683
用語ペア数	1,326	3,225
関係のあるペア数	412	628

表3　用語抽出結果 (従来)

用語	出現回数	F 値
Road	233	0.936
Pass	235	0.881
Car	319	0.900
Driving	378	0.860
Check	164	0.928
その他 (top5 除く)	7503	0.771
全用語	8832	0.789

表4　用語抽出結果 (新)

用語	出現回数	F 値
Driving	389	0.825
Case	736	0.751
Location	793	0.684
Property	848	0.573
Car	323	0.913
その他 (top5 除く)	9243	0.732
全用語	12332	0.726

験は 5 分割交差検証によって行った．新アノテーションでの抽出精度は従来のアノテーション時と比べて高くなった．これは，関係の種類が減ることで予測候補が減り，タスクが簡単になったためと考えられる．特に，range と domain の関係数は大きく増加しており，十分な学習を行うことができ，より精度の高い結果が出たと考えられる．関係ごとの結果も新アノテーションは従来のアノテーションの結果を上回っていた．

アノテーションの変更による用語抽出の精度の相対的な低下が約 10% であるのに対し，関係抽出の精度の相対的な向上は約 40% であり，全体で考えた際の抽出精度の向上が期待できる．

5.2 正誤問題の実験

アノテーションデータを利用して，正誤問題の予測を行った結果を表 7に示す．ベースラインモデルとして，Devlin らが提案した [CLS] のベクトルを分類する文ペアを分類するモデルを用いる．提案手法については，アノテーションの形式を従来のアノテーションのときと新しいオントロジー形式のアノテーションのときでそれぞれ実験を行った．実験は 10 分割交差検証で行い，結果の平均を表 7に示した．従来のアノテーションと新アノテーションのどちらを用いた場合でも提案手法の正答率が最も高かった．また，新アノテーションを用いた場合のほうが高い正答率となっており，提案手法の新アノテーションを用いた場合の結果は実験内の結果の中で最も正答率が高かった．以上のことから新アノテーションを用いた提案手法により，正誤問題の正答率を上げることができることがわかった．

表5 関係抽出結果 (従来)

関係	出現回数	F 値
range	1185	0.481
Location	811	0.383
Case	753	0.298
全関係	7071	0.345

表6 関係抽出結果 (新)

用語	出現回数	F 値
range	4902	0.537
domain	4320	0.496
subClassOf	118	0.367
全用語	10627	0.507

表7 正誤問題の回答精度

	ベース	提案手法	
入力	ライン	従来	新
正答率	0.57	0.60	0.63

6 おわりに

オントロジー形式のアノテーションを用いることで，従来のアノテーションよりも高い精度で関係抽出を行うことができた．これにより，オントロジー形式のアノテーションが知識の抽出を行う上で有用であることがわかった．また，提案したアノテーションを用いることで，正誤問題をより高い精度で判断できることもわかった．今後は，それぞれの手法について，さらなる精度の向上を図るとともに，他の情報抽出タスクへの応用可能性を探る．

謝辞

本研究の一部が JSPS 科研費 17K00318 により支援されたことに深く感謝する。

参考文献

[1] Alan et al. Contextual string embeddings for sequence labeling. In *COLING*, 2018.

[2] Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.

[3] Vaswani et al. Attention is all you need. In *NIPS*, 2017.

[4] 河辺ら．交通オントロジーの半自動拡張のための交通用語認識．言語処理学会第 21 回年次大会発表論文集, 2015.

[5] 国家公安委員会．交通の方法に関する教則．全日本交通安全協会, 2012.

[6] 杉村ら．交通規則問題のための解答システムの構築．言語処理学会第 19 回年次大会, 2013.

[7] 長野．交通オントロジーの階層構造を考慮したニューラル用語抽出．豊田工業大学修士論文, 2018.

[8] 八木ら．CNN を用いた交通文書からの交通用語間関係抽出．言語処理学会第 25 回年次大会発表論文集, 2019.