ドメインを限定した機械読解モデルに基づく述語項構造解析

高橋 憲生　　柴田 知秀　　河原 大輔　　黒橋 禎夫
京都大学 大学院情報学研究科
{ntakahashi, shibata, dk, kuro}@nlp.ist.i.kyoto-u.ac.jp

1　はじめに

これまで、計算機による文章理解を目指して、形態素解析、構文解析、述語項構造解析などの基礎解析研究が活発に行われてきた。現在、日本語の形態素解析の精度は 99%、構文解析の精度は約 90%を越え実用レベルに達しているが、省略現象を含む述語項構造解析の精度は 50%〜60%に留まっている [10][4]。日本語では、主語や目的語が省略されることが多いため、計算機による文章理解を実現するためには、述語項構造解析の精度を向上させる必要がある。

計算機による文章理解という観点では、近年、機械読解研究が盛んに行われている。それらの研究では、文章、質問、答えの 3 つ組からなる QA データセットを構築し、それを用いて機械読解モデルを学習している。機械読解の研究は目覚ましい進歩を遂げており、一部のデータセットでは、人間の精度を越えたという報告 [1] があるが、省略・照応現象が多く含まれる文章、外部知識や推論を必要とする文章の読解については精度はいまだ高くない [5]。

本論文では、特定のドメインに限定した上で、述語項構造解析のデータセットを QA データセット形式で構築し、機械読解モデルに基づく述語項構造解析手法を提案する。また、同じドメインの文章読解データセットも構築し、述語項構造解析データセットと統合的に用いることによる精度向上を目指す。

ドメインは、運転行動に関するブログ記事とする。現代社会において自動車は必要不可欠であるが、あおり運転に代表される危険な運転行動は、安全な交通を阻害している。ブログ記事から運転行動を抽出して、計算機が運転状況を正しく理解できれば、運転状況の安全性の判定や、危険な運転状況の是正を実現でき、安全な交通システムの実現に寄与できる。

運転行動・主観コーパス [11] を用いて、述語項構造と文章読解の 2 種類の QA データセットを構築する。QA データセット構築は、少人数の専門家に依頼せず、大規模かつ短期間でデータセットを作成可能なクラウドソーシングを利用する。機械読解モデルを用いて、述語項構造タスクを単体学習した場合、述語項構造解析と文章読解タスクを統合学習した場合を比較したところ、統合学習の方が高い精度を達成した。

2　関連研究

2.1　QA データセット構築

述語項構造 QA データセットとして、FitzGerald らによって QA-SRL Bank 2.0 が構築されている [2]。これは、英語の意味役割付与に対するアノテーションであり、質問と答え (質問の意味役割に対応する項) をクラウドソーシングで収集している。本研究では、後述するように質問を自動生成し、答えのみをクラウドワーカーに選択させる点で異なる。

文章読解 QA データセットは、近年、多くのデータセットが構築されている。例えば、Rajpurkar らは、論理推論能力を評価するために、ウィキペディア記事から文章の一部を解答する問題を 10 万問作成し、SQuAD 1.1 を構築した [8]。質問に対する答えは、記事中のスパンとなっている。本研究では、SQuAD 1.1 にならい、運転ドメインの文章読解 QA データセットを構築する。

2.2　機械読解モデル

文章読解 QA データセットに対して、ニューラルネットワークを用いた様々な機械読解モデルが提案されている。例えば、Seo らは、文章と質問の間の相互のアテンションを利用する Bi-Directional Attention Flow (BiDAF) Network を提案し、SQuAD 1.1 データセットにおいて高い精度を実現した [9]。

また、複数のデータセットもしくはタスクを対象とした統合的学習手法も提案されている。例えば、Min らは、SQuAD 1.1 の学習結果を転移学習することで、他の機械読解タスクで高い精度を実現した [6]。Pan らは、機械読解タスクで学習した知識を、翻訳タスクや要約タスクに転移学習して、高い精度を実現した [7]。

例 1：述語項構造	例 2：文章読解
・文章：私は右車線に移動した。そしてバックミラーを見た。 ・質問：『見た』の主語は何か？ ・答え：私	・文章：私の車の前をバイクにまたがった警察官が走っていた。 ・質問：警察官は何に乗っていた？ ・答え：バイク

図 1: QA データセットの作成例

3　QA データセットの構築

運転ドメインを対象として、述語項構造 QA データセットと文章読解 QA データセットを構築する。QA データセットの作成には、大規模かつ短期間でデータセットを作成可能なクラウドソーシングを利用する。図 1 の例 1 に述語項構造 QA データセット、例 2 に文章読解 QA データセットの例を示す。

述語項構造 QA データセットは、ガ格、ヲ格及びニ格について省略されている項の先行詞を問う問題である。ガ格省略を問う述語項構造 QA データセットは、以下に示す方法で作成する。

1. 岩井らが構築した運転行動・主観コーパス [11] の中から、以下の条件を満たす 4 文を抽出して、問題用の文章とする。
    - 岩井らによる CRF を用いた自動車運転エピソード抽出ツール [3] が、4 文中 3 文以上自動車運転エピソードと判定する
    - 各文は少なくとも一つの述語項構造を含む
    - 4 文目に、KNP による省略解析が、ガ格が省略されていると判定した述語項構造を含む
    - 4 文の中に、少なくとも一つの「運転特徴語」（[11] 参照）を含む
2. 抽出した問題用の文章と、ガ格省略と判定された述語項構造を基に、図 2 に示すような問題を自動作成する。各問題は、文章、質問、答えの選択肢から構成される。
3. クラウドソーシングを利用して、1 問当たり 5 人のクラウドワーカーに選択肢から回答を選んでもらう。「5 人中 3 人以上の回答が一致」かつ「ワーカーの回答が『その他』及び『分からない』以外」を満たしていれば、人間が解答可能な問題として採用する。

Yahoo!クラウドソーシングを利用して、ステップ 2 で 24,000 問を作成し、最終的に 12,468 問を作成した。構築したガ格省略述語項構造 QA データセットから 200 問を抽出して、ワーカーの回答の種類を確認したところ、60.4%は「著者」であった。また、「著者」に「私」や「僕」などの一人称の答えを含めると、回答の 72.3%は一人称であった。

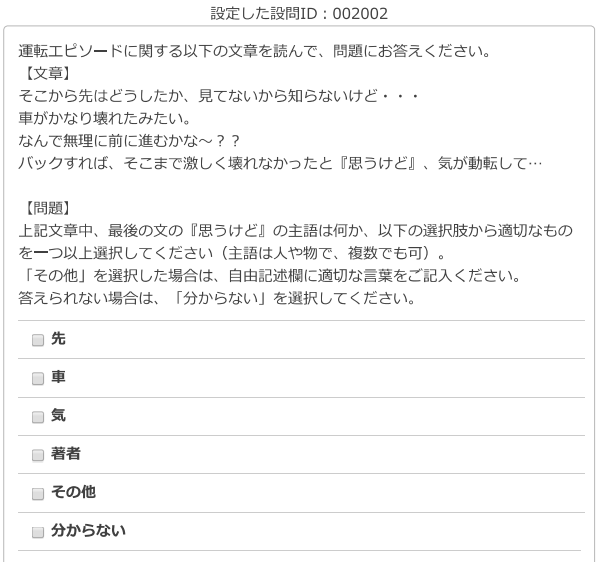

図 2: 述語項構造 QA データセット回答画面の例

	日付	数字	人	場所	名詞句	形容詞句	動詞句
問題数	3	15	31	29	76	6	12
割合	1.6%	7.8%	16.1%	15.1%	39.6%	3.1%	6.3%

表 1: 文章読解 QA データセットの答えの分類

ガ格省略 QA データセットと同様に、ヲ格省略 QA データセットを 1,497 問、ニ格省略 QA データセットを 387 問作成した。

文章読解 QA データセットは、運転行動・主観コーパスから問題用の文章を抽出し、クラウドソーシングで文章から質問と答えを記述してもらう。さらにクラウドソーシングで解答可能な問題か否かを確認する。その結果、20,007 問を作成した。

構築した文章読解 QA データセットの中から 200 問を抽出して、ワーカーの回答の種類を確認した。その結果を表 1 に示す。このように、SQuAD 1.1 と同様に、様々な回答のタイプの問題を作成できた。

また、構築した文章読解 QA データセットの中から 200 問を抽出して、ワーカーの質問の種類を確認した。その結果を表 2 に示す。質問の種類は、ガ・ヲ・ニ格のいずれかの項を問う質問か否か、該当する場合、省略か否かで分類する。表 2 のとおり、文章読解 QA データセットには、ガ・ヲ・ニ格の項を問う質問が 4 割弱含まれており、省略されている項を問う問題も若干含まれている。その他の質問としては、ガ・ヲ・ニ格以外の格の項を問う質問、何故など理由を問う質問、どのくらいなど量を問う質問など様々である。

	ガ格	ヲ格	ニ格	その他
問題数	41	28	8	123
割合	20.5%	14.0%	4.0%	61.5%
（うち省略）	(5.0%)	(2.5%)	(0.5%)	−

表 2: 文章読解 QA データセットの質問の分類

4 機械読解モデルに基づく述語項構造解析

構築した述語項構造 QA データセットを用いて、機械読解モデルに基づく述語項構造解析を行う。述語項構造 QA データセットは、文章、質問、答えの 3 つ組となっているため、既存の高精度な機械読解モデルをそのまま利用することができる。ただし、質問は、図 2 の例のように、クラウドワーカー向けの説明を含んでいるため、機械読解モデルに入力するために、それぞれの格について次のように簡略化する。

- ガ格 :『対象となる述語を含む節』の主語は何か？
- ヲ格 : ○○を『対象となる述語を含む節』、の○○に入るものは何か？
- ニ格 : ○○に『対象となる述語を含む節』、の○○に入るものは何か？

述語項構造 QA データセットは省略されている項のみを対象としているため、本研究における述語項構造解析も同様の項のみが対象となる。また、データセットにおいて常に正解 (先行詞) が存在するため、省略現象の同定は行わず、先行詞同定の問題のみを解くことに相当する。機械読解モデルとしては、文章中のスパンを常に答えとして返すモデルを採用する。

さらに、述語項構造 QA データセットと文章読解 QA データセットの両方を用いた統合学習を行う。統合学習には、両方のデータセットをマージした「同時学習」と、文章読解 QA データセットで学習した後に述語項構造 QA データセットで学習する「逐次学習」の 2 種類を実験する。これによって、文章読解 QA データセットによってドメイン知識が学習され、それが述語項構造解析に効くかどうかを検証する。

5 実験

構築した述語項構造および文章読解 QA データセットを用いて、機械読解モデルに基づく述語項構造解析の実験を行う。まず、述語項構造 QA データセットのみで機械読解モデルを学習する「単体学習」を行う。次に、述語項構造および文章読解 QA データセットの両方を用いた同時学習および逐次学習を行う。

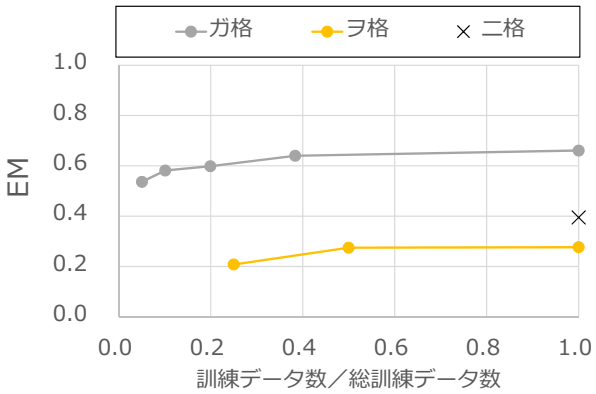

図 3: 述語項構造 QA データセットを用いた単体学習によるラーニングカーブ

5.1 実験設定

機械読解モデルとして BiDAF を用いる。実験には、ガ格省略 QA データセット 12,468 問、ヲ格省略 QA データセット 1,497 問、ニ格省略 QA データセット 387 問及び文章読解 QA データセット 20,007 問を用いる。これらを訓練データ、バリデーションデータ、テストデータに分割する。訓練データ数は、ガ格が 11,359 問、ヲ格が 1,198 問、ニ格が 310 問であり、テストデータ数は、ガ格が 565 問、ヲ格が 149 問、ニ格が 38 問である。各々のデータセットは、Juman++を用いて形態素解析し、前処理を行う。日本語 Web コーパス 2 億文から Word2Vec で学習した単語ベクトルを初期値として学習を行う。精度の評価方法には、EM (Exact Match) を用いる。EM は、（システムが答えた文字列と、データセットの答えの文字列が一致した問題数）／（データセット全体の問題数）である。

5.2 単体学習

図 3 に、機械読解モデル BiDAF による述語項構造 QA データセット単体の解析精度を示す。図の横軸は、学習時に用いた訓練データ数全体に対する訓練データの割合を示す。ガ格及びヲ格については、総訓練データの半分より訓練データを増やしても精度がほとんど上がらなかった。二格については訓練データ数が少ないため、ラーニングカーブを確認していない。

5.3 同時学習及び逐次学習

表 3 に、単体学習、同時学習及び逐次学習のそれぞれの解析精度を示す。

ガ格については、単体学習と同時学習は差が無かったが、逐次学習は 1.6%精度が向上した。ヲ格については、同時学習は単体学習より 4.4%精度が向上した。また、逐次学習は同時学習よりさらに 8.3%精度が向

	ガ格	ヲ格	ニ格
単体学習	0.661	0.277	0.395
同時学習	0.661	0.321	0.395
逐次学習	0.677	0.404	0.493

表 3: 述語項構造 QA データセットの同時学習及び逐次学習結果（数値は EM）

・文章：
　屈伸をしながら気合いを入れ直し坂道に挑む。
　「坂を越えたらバイク屋、坂を越えたらバイク屋」
　他の事を考えないように、それだけをつぶやきながらただただ押す。
　ほんの少し『上っただけで』さっきまで引いていた汗が今まで以上に噴き出し、 坂の真ん中あたりで足がブルブル震え出す。
・質問：
　〇〇を『上っただけで』、の〇〇に入るものは何か？
・答え：
　正解　：**坂**
　単体学習モデル　：汗
　同時学習モデル　：坂道
　逐次学習モデル　：**坂**

図 4: 逐次学習によって正答する例（同時学習の答え「坂道」は本来正解と考えられるが、クラウドソーシングでは「坂」のみが正解となっている）

上した。ニ格については、単体学習と同時学習は差が無かったが、逐次学習は 9.8%精度が向上した。

5.4　議論

図 3 のとおり、ガ格とヲ格については述語項構造 QA データセットを増やしても精度は上がらなかったが、文章読解 QA データセットを加えて統合学習することにより精度が上がった。ヲ格省略 QA データセットについて、単体学習モデル、同時学習モデル及び逐次学習モデルの答えを確認したところ、図 4 の例のように、同時学習モデル及び逐次学習モデルは運転に関する問題の解答精度が向上した。これは、文章読解 QA データセットを学習することにより運転に関する単語やフレーズ間の関係（例えば「上る」と「超える」のような類義関係）を正しく認識することができたためと考えられる。

既存手法との比較も行った。柴田らのニューラルネット省略解析（NN 省略解析）[10] を用いて述語項構造 QA データセットの訓練データを変換して学習したところ、再現率はガ格 0.74、ヲ格 0.30、ニ格 0.07 であった。省略解析は NULL を出力するが、NULL の出力割合は、ガ格がほぼ 0%、ヲ格が 30〜40%、ニ格が 90%以上であった。NN 省略解析の再現率と逐次学習の EM を比較すると、ガ格は NN 省略解析の方が良好、ヲ格とニ格は逐次学習の方が良好であった。

6　まとめと今後の課題

本研究では、運転ドメインに限定した上で、述語項構造 QA データセット及び文章読解 QA データセットを構築した。述語項構造 QA データセットの単体学習より、文章読解 QA データセットを用いた同時学習や逐次学習の方が精度が向上した。今後は、BERT[1] を利用した最新の機械読解モデルなどを採用することによって、述語項構造解析精度をさらに向上させることを目指す。

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, Vol. abs/1810.04805, 2018.

[2] Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. Large-scale QA-SRL parsing. In *ACL2018*, pp. 2051–2060, 2018.

[3] Ritsuko Iwai, Daisuke Kawahara, Takatsune Kumada, and Sadao Kurohashi. Annotating a driving experience corpus with behavior and subjectivity. In *PACLIC 32*, 2018.

[4] Shuhei Kurita, Daisuke Kawahara, and Sadao Kurohashi. Neural adversarial training for semi-supervised Japanese predicate-argument structure analysis. In *ACL2018*, pp. 474–484, 2018.

[5] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *EMNLP2018*, pp. 2381–2391, 2018.

[6] Sewon Min, Min Joon Seo, and Hannaneh Hajishirzi. Question answering through transfer learning from large fine-grained supervision data. In *ACL2017*, pp. 510–517, 2017.

[7] Boyuan Pan, Yazheng Yang, Hao Li, Zhou Zhao, Yueting Zhuang, Deng Cai, and Xiaofei He. MacNet: Transferring knowledge from machine comprehension to Sequence-to-Sequence models. In *NeurIPS2018*, pp. 6095–6105, 2018.

[8] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100, 000+ questions for machine comprehension of text. In *EMNLP2016*, pp. 2383–2392, 2016.

[9] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. *CoRR*, Vol. abs/1611.01603, 2016.

[10] Tomohide Shibata and Sadao Kurohashi. Entity-centric joint modeling of Japanese coreference resolution and predicate argument structure analysis. In *ACL2018*, pp. 579–589, 2018.

[11] 岩井律子, 熊田孝恒, 高橋憲生, 河原大輔, 黒橋禎夫. 心理表現を含む運転関連表現辞書の構築. 言語処理学会第 25 回年次大会論文集, 2019.