一般セッション | 一般セッション | [GS] J-9 自然言語処理・情報検索

[1N4-J-9] 自然言語処理・情報検索: ドメイン知識分析

座長:大熊 智子(富士ゼロックス)　評者:貞光 九月(フューチャー株式会社)

2019年6月4日(火) 17:20 〜 18:40 N会場 (1F 展示ホール右奥)

[1N4-J-9-01] 自己学習による化学文書中の専門用語抽出

〇崔 一鳴[1,3]、西川 仁[1,3]、徳永 健伸[1]、吉川 和[2,3]、岩倉 友哉[2,3] （1. 東京工業大学 情報理工学院、2. 株式会社富士通研究所、3. 理研 AIP-富士通連携センター）

17:20 〜　17:40

[1N4-J-9-02] 複数サブワード系列を考慮した BiLSTM-CRFモデルを用いた文書からの化合物名抽出

〇関根 裕人[1]、浦澤 合[1]、乾 孝司[1]、岩倉 友哉[2] （1. 筑波大学院/理研 AIP-富士通連携センター、2. 理研 AIP-富士通連携センター）

17:40 〜　18:00

[1N4-J-9-03] 文書からの化合物名抽出のためのサフ゛ワート゛有効性調査

〇浦澤 合[1,3]、関根 裕人[1,3]、乾 孝司[1,3]、岩倉 友哉[2,3] （1. 筑波大学大学院、2. 富士通研究所、3. 理研 AIP-富士通連携センター）

18:00 〜　18:20

[1N4-J-9-04] 運転免許試験で使用される語彙と省略語句の分析

〇的場 成紀[1]、古賀 雅樹[1]、大塚 基広[1]、小林 一郎[2]、平 博順[1] （1. 大阪工業大学大学院 情報科学研究科、2. お茶の水女子大学大学院 人間文化創成科学研究科）

18:20 〜　18:40

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

自己学習による化学文書中の専門用語抽出

Chemical Named Entity Recognition with Self-Training

崔一鳴 *1*3 Yiming Cui　西川仁 *1*3 Hitoshi Nishikawa　徳永健伸 *1 Takenobu Tokunaga　吉川和 *2*3 Hiyori Yoshikawa　岩倉友哉 *2*3 Tomoya Iwakura

*1東京工業大学 情報理工学院 School of Computing, Tokyo Institute of Technology　*2株式会社富士通研究所 Fujitsu Laboratories Ltd.

*3理研 AIP-富士通連携センター RIKEN AIP-Fujitsu Collaboration Center

In this paper, we propose to use self-training for chemical named entity recognition. We first train a neural network-based model for chemical named entity recognition model using the CHEMDNER corpus. The trained model is used to annotate the unlabelled MEDLINE corpus to create automatically labelled training data. We then use both training data, manually labelled CHEMDNER corpus and automatically labelled MEDLINE corpus, to train our final model. The evaluation using the unlabelled MEDLINE corpus as training data showed that the effectiveness of self-training in the chemical named entity recognition task.

1. はじめに

化学分野の研究は非常に盛んであり，日々新たな発見がなされ，論文が発表されている．それらの論文の中には今まで登場したことのない，新しい化学用語が出現する．化学用語のデータベースは化学分野の研究において重要な言語資源であるが，現状ではそのデータベースは新しく出版された論文や特許を人手によって読解し，新しく出現した用語を抽出し構築されている．この作業は費用を要する作業であるだけでなく，時間的な面においても困難であり，新しい用語の登場の速度に人手による抽出によって追随することは容易ではない．さらに，化学分野に精通した人員でなければ新しい化学用語の抽出は難しい作業であるため、人員の確保そのものも容易ではない．そのため，論文からの情報抽出の自動化は急務である．

この問題を解決するため，化学文書からの化学に関する専門用語の自動抽出を試みる研究が進められている [14]．化学用語の自動抽出課題のために作られたコーパスである CHEMDNER コーパス [4] においては，10,000 件の化学分野の論文のアブストラクト（訓練データ 3,500 件，開発データ 3,500 件，テストデータ 3,000 件）に対して，化学用語部分にラベルが付与されている．このコーパスにおいて現時点での最高精度を報告している研究は Bi-LSMT-CRF[5] に注意機構を加えたモデルであり，F 値 91.14%を達成している [6]．同じデータに対して，専門家が化学用語の抽出を行った結果の一致率は 89%であるため，それを上回る精度を達成したことになる．

本研究では，この化学用語抽出課題の精度をさらに向上させる試みとして，自己学習によって CHEMDNER 以外の大規模なデータを利用する手法を提案する．具体的には，CHEMDNER の訓練データを用いて作成したモデルを用いて，化学用語部分にラベルが付与されていない MEDLINE コーパス [11] のアブストラクトにラベルを付与し，それを新しい訓練データとして，化学用語抽出のモデルを作成する．

連絡先:
*1{sai.m.ab@m, hitoshi@c, take@c}.titech.ac.jp
*2{y.hiyori, iwakura.tomoya}@jp.fujitsu.com

2. 関連研究

化学用語の抽出については数多くの研究がなされており，類似する課題である固有表現抽出 [2] と同様に，機械学習を用いる手法が数多く提案されている．近年は，ニューラルネットワークを用いる手法が活発に研究されており， CHEMDNER コーパスにおける現時点での最高の性能を報告している論文 [6] は双方向 LSTM （ Bi-LSMT ）と条件付き確率場（ CRF ）を組み合わせた Bi-LSMT-CRF[3, 7] に基づく．先行研究に倣い，本研究でも Bi-LSMT-CRF をベースラインとして学習を行い，モデルを作成する．Luo らは Bi-LSMT-CRF に注意機構 [1] を加えたモデルを採用し，最高精度を報告している．Luo らは入力の特徴量として，MEDLINE コーパスと CHEMDNER コーパスで word2vec を用いて単語分散表現を用いている．また，品詞データや Bi-LSMT による文字分散表現 [12] も入力の特徴量として用いている．

自己学習は自然言語処理において広く用いられており，構文解析 [8] や語義曖昧性解消 [9] などに利用されている．

ニューラルネットを利用したモデルに自己学習を利用した先行研究として竹前らが見出し生成課題に自己学習を利用したもの [16] がある．竹前らは，正例が付与されているデータからまずモデルを作成し，それを正例が付与されていないデータに対して適用することで疑似的な正例データを作成した．その上で正例が付与されているデータと，疑似的な正例データの両者を用いて最終的なモデルを作成した．自己学習を行ったモデルと行っていないモデルを比較することによって，竹前らは見出し生成課題において自己学習を行うことによって性能が向上することを示した．

本論文は，化学用語抽出課題において，自己学習を行うことを提案する．まず化学用語ラベルが付与されているデータセットである CHEMDNER コーパスを用いてモデルを学習し，化学用語ラベルが付与されていない MEDLINE コーパスに疑似的な正解ラベルを付与する．その後， CHEMDNER コーパスと疑似的な正解ラベルが付与された MEDLINE コーパスのデータを併せて再度モデルを作成し，これを最終的なモデルとする．

3. 提案手法

本論文の提案手法は以下のような手順になる．

1. 教師あり学習：まず， CHEMDNER の訓練データを利用して教師あり学習を行う．これをベースラインモデルとする．
2. 疑似教師データの作成：次に化学用語のラベルが付与されていない，テキストのみのデータに対して手順 1 で作成したモデルを用いて，化学用語ラベルを付与する．これを疑似訓練データとする．
3. 新規モデルの学習：手順 1 で利用した CHEMDNER の訓練データと，手順 2 で化学用語ラベルを付与した疑似訓練データの両方を用いて，新しいモデルの学習を行う．その後， CHEMDNER のテストデータを用いて，このモデルの精度の評価を行う．
4. 手順 2 と手順 3 を繰り返し，最終的なモデルを得る：手順 3 で得られたモデルの性能の評価が手順 1 で得られたモデルよりも高精度である場合，手順 3 で得られたモデルを再度用いて手順 2 と手順 3 を行う．精度向上がみられなかった場合学習を終了し，最終的なモデルを得る．

4. 実験

4.1 ベースラインモデル

ベースラインのモデルは，単語分散表現と Bi-LSMT-CRF を利用している．構成図を図 1に示す.

4.1.1 単語分散表現

単語系列をモデルに入力する際には， word embeddings 層を通して，単語を単語分散表現に変換している．その際には gensim[*1] による word2vec[10] を利用した．使用データは CHEMDNER コーパスの訓練データおよび単語分散表現を得る際には，スペースで区切られているものを 1 単語とし， パラメータとして次元数は 100， window size は 4，iter は 10 とした．また min count は 0 とし，学習時に登場した単語を全て分散表現の辞書に登録した．

加えて，文字の分散表現の情報も gensim ライブラリの word2vec を用いて獲得した．文字分散表現は 10 次元とし、こちらも CHEMDNER コーパスの訓練データおよび MEDLINE コーパスを利用した。

最終的は，ある単語の単語分散表現は，単語そのものの分散表現 100 次元と，単語を構成する文字の分散表現 10 次元の合計 110 次元となる．文字の分散表現は，単語に含まれている文字分散表現の平均を取った 10 次元とした．

4.1.2 Bi-LSMT-CRF

本研究におけるベースラインのモデルは Bi-LSMT-CRF によって構築されている．これは PyTorch[*2] を用いて実装を行なった。LSTM の隠れ層は 150 次元とした．また，テストの際に，分散表現辞書にない未知語が現れた際には，ランダムなベクトルを生成し，割り当てた．登場する単語が未知でも，その単語に含まれている文字情報は未知ではないことが多い．単語が未知の場合は，ランダムに生成した 100 次元のベクトルと，文字情報から得られるランダムではない 10 次元のベクトルを合成して，110 次元のベクトルとした．また，本研究では未知語にランダムなベクトルを割り当てる際に，同じベクトルが割り当てられるよう seed 値を固定して乱数を生成した．

4.1.3 化学用語ラベル

化学用語ラベルには固有表現抽出課題で広く利用されている IOB2[13] タグ方式を採用した．これは，化学用語を構成する最初の単語に B，最初の単語ではないが化学用語の一部である単語には I，化学用語ではない単語には O というタグを付与するというものである．

4.1.4 学習

このモデルを CHEMDNER コーパスの訓練データを用いて学習を行ったものをベースラインのモデルとした．学習の際のにはミニバッチ学習を行い、ミニバッチのサイズは 100 とした．また，epoch 数は最大 20 とした．

4.2 比較手法

- ベースライン手法： CHEMDNER コーパスの訓練データを用いて，前述のモデルを用いた化学用語抽出モデルを学習し，これをベースラインの手法として利用した．
- 提案手法：ベースラインの手法によって学習が行われたモデルを用いて、化学用語ラベルが付与されていない MEDLINE コーパスのデータに化学用語ラベルを付与し，それを疑似訓練データとし，前述したように CHEMDNER の訓練データと MEDLINE 疑似訓練データの両者を利用することで化学用語抽出モデルを構築する．

4.3 データ

4.3.1 CHEMDNER コーパス

CHEMDNER コーパスは化学用語抽出課題のために構築されたコーパスである [4]．コーパスの統計量を表 1に示す．CHEMDNER コーパスは化学関連論文のアブストラクト 10,000 件からなり，化学用語の箇所に化学用語ラベルが付与されている．化学用語は， TRIVAL， SYSTMATIC， ABBREVIATION， FORMULA， FAMILY， IDENTIFIER, MULTIPLE， NO CLASS の 8 クラスに分類されラベルが付与されているが，本実験では先行研究 [6] と同様に，これら 8 クラスを等しく化学用語であるとして，すなわち単一のクラスとして扱う．

4.3.2 MEDLINE コーパス

MEDLINE は医学を中心とする文献情報を収集したオンラインデータベースである．[15] このデータベースは医学，薬学，看護学，歯学，衛生学，獣医学，生化学、分子生物学など医学に関連する幅広い文献情報を含んでいるが，本研究では，MEDLINE コーパスの中でも，2017 年版の CHEMDNER コーパス作成時に対象としたジャーナル・会議のもののみを利用した．MEDLINE コーパスには，CHEMDNER コーパスとは異なり，化学用語ラベルは付与されておらず，テキストのみのデータのみとなっている．使用した MEDLINE コーパスに含まれるアブストラクトの文字数の合計は約 18 億字であり，これは CHEMDNER コーパスの訓練データ約 488 万字の 375 倍の量に相当する．また，単語数は約 2 億 7 千万語あり，これは CHEMDNER コーパスの訓練データ約 77 万語の約 350 倍に相当する．前述したように，本研究ではこの MEDLINE コーパスに対して CHEMDNER コーパスで訓練したモデルを利用して化学用語ラベルの付与を行い，これを疑似訓練データとして利用する．

4.4 評価

学習を終えた化学用語抽出モデルは CHEMDNER コーパスのテストデータを利用して行う．また，本研究が対象とする

*1 https://radimrehurek.com/gensim/
*2 https://pytorch.org/

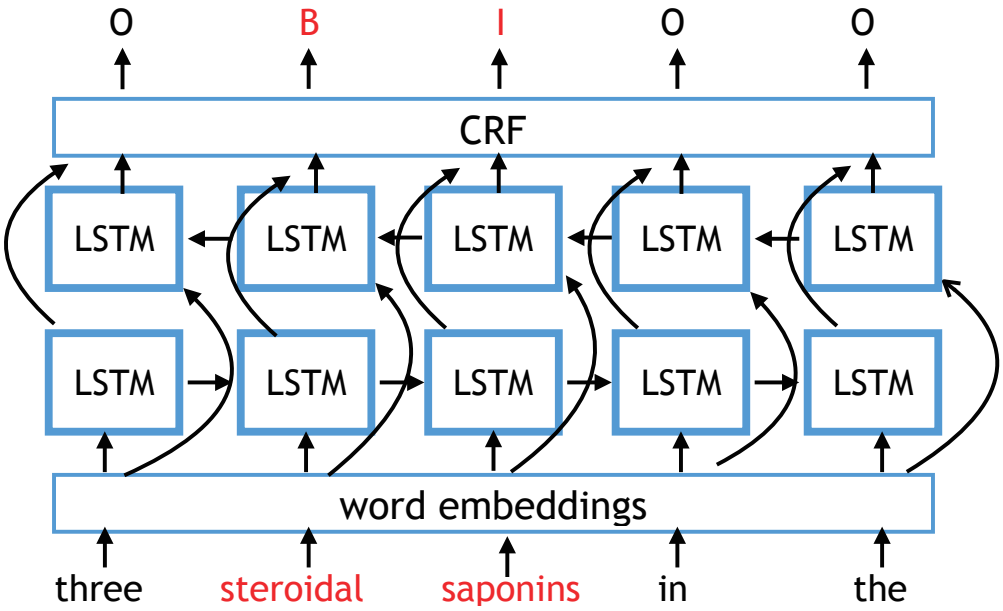

図 1: 実装したモデルの概念図

表 1: CHEMDNER コーパスの基本データ

	訓練データ	開発データ	テストデータ	合計
アブストラクト数	3,500	3,500	3,000	10,000
全文字数	4,883,753	4,864,558	4,199,068	13,947,379
全単語数	770,855	766,331	662,571	2,199,757
TRIVIAL	8,832	8,970	7,808	25,610
SYSTEMATIC	6,656	6,816	5,666	19,138
ABBREVIATION	4,538	4,521	4059	13,118
FORMULA	4,448	4,137	3,443	12,028
FAMILY	4,090	4,223	3,622	11,935
IDENTIFIER	672	639	513	1,824
MULTIPLE	202	188	199	589
NO CLASS	40	32	41	113

表 2: 実験結果

使用データ	精度	再現率	F 値
ベースライン	0.866	0.787	0.824
自己学習 1 回	0.867	0.812	0.839
自己学習 2 回	0.857	0.826	0.841
自己学習 3 回	0.842	0.837	0.839

化学用語抽出課題と同様に線形ラベル付け問題である固有表現抽出課題においては，一般的に精度，再現率，および F 値が評価尺度として用いられるため，本研究においてもこれらの値を評価尺度として用いる．

5. 結果と考察

CHEMDNER コーパスのテストデータを用いて評価した結果を表 2に示す．ベースラインモデルに比べて，自己学習を行うことで F 値が向上していることがわかる．特に，提案手法の手順を繰り返すことによって再現率が繰り返し向上しており，このことは訓練データの網羅性の不足が自己学習によって補われていることを示唆している．一方で，精度は自己学習を 1 回だけ行った場合が最もよい．これは自己学習によって本来は正しくない単語列が化学用語として疑似訓練データに混入することが増えることによって，テストデータにおける精度が低下するためと思われる．結果としては自己学習を 2 回行った際の F 値が最も良好な結果を得た．

また，手順 2，3 で用いるデータの量を変化させた際の精度の変化を表したグラフを表 3に示す．この結果は自己学習を 1 回だけ行った際の結果である．追加データの量は，CHEMDNER の訓練データの量を 100%とした際の疑似訓練データの量を示す．すなわち，0%は自己学習を行わない場合（表 2のベースライン）に相当する．全体的に，疑似訓練データを増やすことで，ベースラインのモデルと比較して F 値が向上することがわかる．結果として 450%において最高精度を記録しているものの，その一方で，データ量を増加させたからといって一貫して F 値が向上するという傾向は見られず，データを追加したからといって安定して性能が向上するとは言えない．そのため，自己学習に利用するデータを増加させることによって安定して性能を向上させるためには何らかの追加的な工夫が必要となるものと思われる．

表 3: データ量ごとの結果

追加データの量	精度	再現率	F 値
0 %	0.866	0.787	0.824
50 %	0.879	0.789	0.832
100 %	0.870	0.808	0.838
150 %	0.860	0.812	0.838
200 %	0.867	0.812	0.839
250 %	0.877	0.800	0.836
300 %	0.849	0.832	0.841
350 %	0.877	0.800	0.836
400 %	0.875	0.801	0.836
450 %	0.883	0.804	0.842
500 %	0.860	0.822	0.841
550 %	0.883	0.804	0.837

6. まとめ

本論文では，Bi-LSMT-CRF による化学用語抽出のモデルを学習する際に，自己学習を利用してラベルの付与されていないデータを利用することを提案した．実験により，自己学習を利用することでモデルの性能が向上したことが示された．また，自己学習を利用して得られたモデルで再度ラベルの付与を行い，それを利用して再び自己学習を行うことで，さらに性能が向上することも確認した．自己学習の際に利用するデータを増加させることによって F 値が向上することも確認できたものの，F 値の向上は一定しておらず，データ量を増加させる際には何らかの工夫が必要であると思われる．本研究は最高精度に到達していないが，現在最高精度を出しているモデルを再現し自己学習を加えることで，最高精度に到達できる見込みがあるものと思われる．

参考文献

[1] Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1462–1472, 2016.

[2] Asif Ekbal and Sivaji Bandyopadhyay. Named entity recognition using support vector machine: A language independent approach. *International Journal of Electrical, Computer, and Systems Engineering*, Vol. 4, No. 2, pp. 155–170, 2010.

[3] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[4] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, Roger A Sayle, Riza Theresa Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktschel, Srgio Matos, David Campos, Buzhou Tang, Hua Xu, Tsendsuren Munkhdalai, Keun Ho Ryu, SV Ramanan, Senthil Nathan, Slavko itnik, Marko Bajec, Lutz Weber, Matthias Irmer, Saber A Akhondi, Jan A Kors, Shuo Xu, Xin An, Utpal Kumar Sikdar, Asif Ekbal, Masaharu Yoshioka, Thaer M Dieb, Miji Choi, Karin Verspoor, Madian Khabsa, C Lee Giles, Hongfang Liu, Komandur Elayavilli Ravikumar, Andre Lamurias, Francisco M Couto, Hong-Jie Dai, Richard Tzong-Han Tsai, Caglar Ata, Tolga Can, Anabel Usi, Rui Alves, Isabel Segura-Bedmar, Paloma Martnez, Julen Oyarzabal, and Alfonso Valencia. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, Vol. 7, No. Suppl 1, pp. 1–17, 2015.

[5] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

[6] Ling Luo, Zhihao Yang, Pei Yang, Zhang Yin, Lei Wang, Hongfei Lin, and Jian Wang. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics*, Vol. 34, No. 8, pp. 1381–1388, 2018.

[7] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

[8] David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics*, pp. 152–159. Association for Computational Linguistics, 2006.

[9] Rada Mihalcea. Co-training and self-training for word sense disambiguation. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, 2004.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[11] U.S. National Library of Medicine. Medline. https://www.nlm.nih.gov/databases/download/pubmed_medline.html.

[12] Marek Rei, Gamal KO Crichton, and Sampo Pyysalo. Attending to characters in neural sequence labeling models. *arXiv preprint arXiv:1611.04361*, 2016.

[13] Erik F Sang and Jorn Veenstra. Representing text chunks. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pp. 173–179. Association for Computational Linguistics, 1999.

[14] Miguel Vazquez, Martin Krallinger, Florian Leitner, and Alfonso Valencia. Text mining for drugs and chemical compounds: methods, tools and applications. *Molecular Informatics*, Vol. 30, No. 6-7, pp. 506–519, 2011.

[15] Beatriz Vincent, Maurice Vincent, and Carlos Gil Ferreira. Making pubmed searching simple: learning to retrieve medical literature through interactive problem solving. *The oncologist*, Vol. 11, No. 3, pp. 243–251, 2006.

[16] 竹前慎太郎, 村尾一真, 谷塚太一, 小林隼人, 野口正樹, 西川仁, 徳永健伸. 自己学習を用いたニューラル見出し生成. 人工知能学会全国大会論文集 2018 年度人工知能学会全国大会 (第 32 回) 論文集, pp. 3Pin136–3Pin136. 一般社団法人 人工知能学会, 2018.

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

複数サブワード系列を考慮した BiLSTM-CRF モデルを用いた文書からの化合物名抽出

関根裕人 *1 浦澤合 *1 乾孝司 *1 岩倉友哉 *2
Hiroto Sekine Go Urasawa Takashi Inui Tomoya Iwakura

*1筑波大学大学院/理研 AIP-富士通連携センター
*2富士通研究所/理研 AIP-富士通連携センター

In this paper, we propose a BiLSTM-CRF model for extracting compound names from documents in chemical domain. The proposed model can be taken multiple subword sequences as input in order to obtain sufficient features for long span or unknown tokens. Subword LSTM units with contextual information are introduced in the input layer of the model. We conducted experiments based on CHEMDNER challenge to investigate the effectiveness of the model. As a result, the extraction accuracy outperformed the normal BiLSTM-CRF model, and experimental results on unknown words showed that the proposed method works better.

1. はじめに

1.1 研究背景

人間が読める言語で書かれた文書から自動的に構造化データを抽出するタスクを「情報抽出」という。近年、Bioinfomarmatics 分野では、この情報抽出を利用した、化合物名抽出という研究分野がある。テキストから化合物名を抽出しデータを構造化することで、論文の検索性を向上させたり、データを分析し化合物間の関係について分析することができる。例えば、化合物名抽出の競争的イベントの一つに CHEMDNER[1] がある。このタスクでは PubMed という化合物関連の論文サイトの論文をアノテーションしたコーパスを作成し、コーパスとして公開している。そして、このコーパスを使用した研究活動が活発に行われている。

上記の CHEMDNER で扱われる化合物名抽出には、一般的な固有表現抽出よりも難しい点がいくつか存在する。化合物名は一つのエンティティの長さが長いものがある。例えば、“3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide”は抽出すべき一つのエンティティであるが、このエンティティは全部で 58 文字から構成されている。次に、同一のエンティティでも複数の名前がつけられていることである。フェニルアラニンがその一例である。フェニルアラニンの他に、“L-Phenylalanine”,“Phe-OH”,“Antibiotic FN-1636 ”など 30 種類以上の呼び名が存在する。また、未知語が多いということも難しい点である。化合物は日々新しく作られていて、それに伴い化合物名も増加していく。未知語に対してどの程度抽出できるかという点も重要な評価の要素となる。

1.2 目的

近年、固有表現抽出の分野では Long Short Term Memory Network(LSTM) を用いて抽出されることが多い。LSTM を使用すると単語の意味や文脈の関係を考慮した計算が可能となる。本稿では LSTM をベースとして、上記の問題を解決するためにサブワードの情報をモデルに組み込む手法を提案する。サブワードとは一つの単語をより細かい単位に分ける考え方である。化合物名には極端に長い単語があるので、サブワード化により細かく区切ることで単語のみの時よりも高い抽出精度が達成できると期待できる。

連絡先: 関根裕人，筑波大学院システム情報工学科，sekine@mibel.cs.tsukuba.ac.jp

1.3 本論文の構成

本稿ではまず、第 2 章で LSTM による固有表現抽出の関連研究をいくつかあげる。第 3 章ではベースラインとなる BiLSTM-CRF モデルについて説明する。第 4 章ではサブワードの分散表現の獲得手法について述べる。そして第 5 章で実験とその結果について述べる。最後に第 6 章でまとめを行う。

2. 関連研究

近年の化合物名抽出タスクでは固有表現抽出を系列ラベリング問題に落とし込み、ニューラルネットワークを用いて解く手法が中心である。その中でも Long Short Term Memory(LSTM) を用いたものが多い。

Jie[2] らは Neural Network による系列ラベリングの手法をまとめ、その抽出精度を測った。比較対象は LSTM や CNN などのモデルによる違いや、学習パラメータの違いや、文字系列の有無による違いなどをまとめている。この実験では固有表現抽出に対しては、BiLSTM-CRF に文字系列 LSTM を加えた結果が最も値がよかった。

Luo ら [3] は既存の BiLSTM-CRF に加えて、Attention 層を追加したモデルを提案した。Attention 層では、global vector という文全体の類似度を考慮することで、より幅の広い時系列の情報も取り込むことができるようになっている。

Akbik[4] らは BiLSTM-CRF モデルに加えて文字系列 LSTM を拡張した。一般的に文字系列の LSTM は一つの単語内に適用されるが、文全体の中で文字系列 LSTM を計算することによって、文脈を残しながら文字をベクトルに埋め込むことができると提案している。

このように単語および文字の情報を考慮した手法は多く提案されているが、本研究のようにサブワードを考慮した手法は提案されてない。

3. ベースライン手法

3.1 BiLSTM-CRF モデル

BiLSTM-CRF モデルについて説明する。モデルの全体構成を図 1 に示す。入力する単語系列、ラベル系列をそれぞれ $x = (x_1, x_2, ..., x_t, ..., x_N)$、$y = (y_1, y_2, ..., y_t, ..., y_N)$ とする。x_t は t 番目の単語の分散表現、y_t はその単語に対応するラベルを表している。

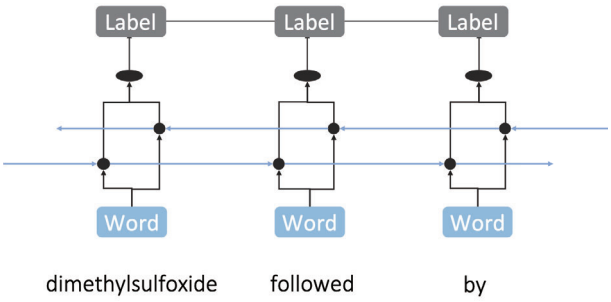

図 1: BiLSTM-CRF モデルの全体図

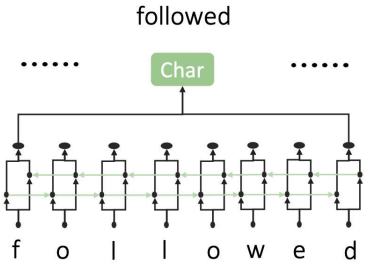

図 2: 文字 BiLSTM モデル

モデルはまず与えられた入力に対し、LSTM 層で計算を行う。LSTM とは時系列ニューラルネットワーク (RNN) の一種である。一つ前のステップの出力に加えて、時系列内で重要と考えられる情報をゲート構造を用いて保持する長期記憶と呼ばれる隠れ層を保持している。これにより一般的な RNN よりも長期的な依存関係を考慮して計算している。

$$h_t, c_t = f_{LSTM}(x_t, h_{t-1}, c_{t-1}; \theta) \quad (1)$$

BiLSTM-CRF モデルでは LSTM を前と後ろからの両方向から計算する。前向きの LSTM の出力を $\overrightarrow{h}$、後向きの LSTM の出力を $\overleftarrow{h}$ とする。これらの計算結果を各ステップごとにつなぎ合わせ、活性化関数の tanh をかける。

$$out = \tanh([\overrightarrow{h} \oplus \overleftarrow{h}]) \quad (2)$$

最後に CRF 層で、それぞれのラベル列の遷移確率を考慮し、入力系列 x に対してもっともらしい y を求める。学習時は以下の $P(y|x)$ を最大にするように、パラメータを更新する。

$$P(y|x) = softmax(Score(x, y)) \quad (3)$$

この時の $Score$ 関数は、各ラベルごとの遷移確率を $T[y_{t-1} \to y_t]$ とすると、以下の式で表すことができる。

$$Score(x, y) = \sum_{t=1}^{N} (\log(out_t) + \log(T[y_{t-1} \to y_t])) \quad (4)$$

3.2 Character Representation

単語系列に加えて、文字系列の分散表現を BiLSTM-CRF モデルに追加することで、抽出精度が上がることが多い。特に未知語に対しては、単語では情報がなくなってしまう場合でも、文字を使用することで情報を得ることができる。今回は文字系列の埋め込みを LSTM を使用して獲得する。

図 2 の部分が文字系列 LSTM の概要である。一つの単語を文字に分解し、文字単位で BiLSTM 層で計算する。前向き LSTM では一番最後の隠れ層の出力、反対に後向き LSTM では一番先頭の出力をそれぞれ連結させる。 最終的に、このベクトルと単語の分散表現に連結して、BiLSTM-CRF モデルの入力として使用する。

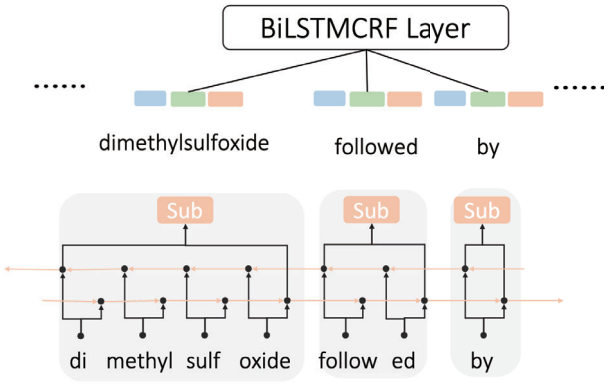

図 3: SubWord LSTM モデルの全体図

4. Subword Representation

サブワードの分散表現の獲得方法について述べる。サブワードとは、単語をさらに分割することでより意味のある情報を得ようとする考え方である。例えば“dimethylsulfoxid”という語は“di, methyl, sulf, oxid”という分割をすると“2”を意味する“di”や、“メチル基”を表す“methyl”、“酸”を表す“oxid”などの情報を得ることができる。

4.1 Subword LSTM

サブワードの分散表現を獲得するための新しいモデルを提案する。このモデルの全体図を図 3 に示す。本手法は Akbik[4] から着想を得て、サブワードを考慮するように変形しものである。

ある単語 x_t が m 個のサブワードに分割された時、$x_t = s_{x_t,1}, s_{x_t,2}, ..., s_{x_t,m_t}$ と表す。このとき、入力 $x = (x_1, x_2, ..., x_N)$ から得られるサブワード系列は $S = (s_{x_1,1}, ..., s_{x_1,m_1}, s_{x_2,1}, ..., s_{x_N,m_N})$ と表せる。

得られた S に対し、第 3 章で述べた BiLSTM 層と同様に計算する。今回、サブワード系列のベクトルを Word 系列に合わせる必要がある。前向き LSTM の計算結果では単語の最後のサブワード、後向き LSTM の計算結果では単語の先頭のサブワードに対応するベクトルを選択する。

ここで、前向き LSTM の出力を $\overrightarrow{h_t} = f_{LSTM}(s_t)$ とする。これを Word 系列と同じ長さにするためには、$(s_{x_1,m_1}, s_{x_2,m_2}, ..., s_{x_N,m_N})$ と対応する $\overrightarrow{h_t}$ を前向き LSTM の計算結果 $\overrightarrow{h_w}$ として使用する。

反対に、後向き LSTM を Word 系列に連結するためには、$(s_{x_1,1}, s_{x_2,1}, ..., s_{x_N,1})$ と対応する $\overleftarrow{h_t}$ を後向き LSTM の計算結果 $\overleftarrow{h_w}$ として使用する。

表 1: 実験結果

	Precision	Recall	Fscore
BaseLine(82ep)	0.9031	0.8578	0.8799
+ SW2k(93ep)	0.9047	0.8584	0.8809
+ SW4k(82ep)	0.9032	0.8589	0.8805
+ SW16k(57ep)	0.8998	0.8577	0.8783
+ SW4k,16k(86ep)	0.9006	0.8668	0.8834
+ SW2k, 4k,16k(80ep)	0.9025	0.8566	0.8790

表 2: IV に対する実験結果

	Precision	Recall	Fscore
BaseLine(82ep)	0.9095	0.8941	0.9018
+ SW2k(93ep)	0.9127	0.8921	0.9020
+ SW4k(82ep)	0.9102	0.8921	0.9011
+ SW16k(57ep)	0.9089	0.8917	0.9002
+ SW4k,16k(86ep)	0.9097	0.8976	0.9036
+ SW2k, 4k,16k(80ep)	0.9090	0.8940	0.9014

表 3: OOV に対する実験結果

	Precision	Recall	Fscore
BaseLine(82ep)	0.8663	0.6871	0.7664
+ SW2k(93ep)	0.8628	0.6939	0.7692
+ SW4k(82ep)	0.8630	0.6894	0.7665
+ SW16k(57ep)	0.8619	0.6876	0.7649
+ SW4k,16k(86ep)	0.8635	0.7065	0.7771
+ SW2k, 4k,16k(80ep)	0.8507	0.6995	0.7677

最後に、前向き LSTM と後向き LSTM の計算結果をつなぎ合わせた $[\overrightarrow{h_w} \oplus \overleftarrow{h_w}]$ がサブワードの分散表現となる。このベクトルは、文字の分散表現と同様に、BiLSTM-CRF モデルへの入力に連結されて使用される。

このモデルは複数のサブワード系列を考慮する場合でも使用できる。その場合は、Subword LSTM をサブワード系列の数だけ用意し、一つの場合と同様に分散表現を得て、BiLSTM-CRF モデルへの入力として、サブワード系列の数だけベクトルを連結し使用する。

5. 評価実験

5.1 実験内容

提案モデルの有効性を調査する。ベースラインとして BiLSTM-CRF に文字 LSTM を加えたものを使用する。このベースラインに Subword LSTM を加えたときの抽出精度の有効性を調査する。

5.2 データ

BioCreative Challenge から出された Chemdner コーパス [1] を実験用のデータとする。このコーパスは PubMed 中の論文の abstract を 10,000 件集め、それらに化合物と判断したエンティティを人手でアノテーションしたものである。全部で 84,355 のエンティティが存在し、それらのユニークな数は 19,806 である。データ数は訓練、検証、テスト用それぞれ 3,500、3,500、 3,000 件ずつ提供されている。また本研究では、BIOES スキーマに従ってラベルづけを行った。

事前学習用のコーパスとして化学系の論文を扱うサイトである PubMed から Chemdner タスクに合うように選択された約 440 万件の abstract を使用した。単語系列、サブワード系列それぞれの埋め込み層の学習には GloVe[5] を使用した。

また、このコーパスを使用して SentencePiece[6] の学習も行った。学習にはユニグラムを使用し、語彙数は 2,000、4,000、16,000 の 3 つを使用した。

5.3 実験パラメータ

今回の実験では最適化に SGD を使用し、学習率は 0.005、減衰率は 0.0001 とした。単語系列、文字系列、サブワード系列の分散表現の次元はそれぞれ、50 次元、30 次元、50 次元とし、LSTM の隠れ層では 200 次元、50 次元、50 次元とした。

また、GPU は Tesla V100-DGXS を用いて学習した。バッチサイズが 10 で、100 エポック学習させた時に検証用データに対してもっとも値の良いモデルを使用した。

5.4 実験結果

実験結果を表 1 にまとめた。語彙数が 2000、4000 のサブワード系列をそれぞれひとつずつ加えた場合、ベースラインよりも F 値を上回った。これより、ベースラインにサブワード系列を追加した場合、抽出精度が向上することがある。

今回の実験で最も良い F 値であったのは、語彙数が 4,000、16,000 のサブワード系列を同時に加えた場合であった。ベースラインと比較すると、0.003 上回っており、0.8834 であった。これは、複数のサブワード系列を同時に加えることで、抽出精度が良くなる場合があることを示している。

サブワードの有効性を調べるために未知語の単語に対する結果も調査した。抽出すべきエンティティが全て未知語だった場合のエンティティを OOV(Out of vocaburaly) と呼び、反対に未知語以外の語が一つでも入っている語を IV(In vocaburaly) と呼ぶ。テストデータの全エンティティ25,308 個に対して、IV は 20,859 個で、OOV は 4,449 個であった。OOV の例としては“HANPs”, “nitriles”, “fidarestat”, “flavonolignans”, “SFN”, “CDDP”, “CYN” などがあげられる。

結果を表 2 および表 3 に示す。表から、IV に対しての結果はあまり変化がみられなかった。しかし、OOV に対してはサブワードを追加したモデルがベースラインよりも Recall が上回っていた。これは、未知語に対してサブワードの埋め込みベクトルがうまくはたらいていることを示している。また OOV のときでは、SW4k,16k はベースラインよりも 0.01 以上も F 値を上回っていた。OOV において、ベースラインでは抽出できなかったが、SW4k,16k で抽出できた例としては、“HANPs”、“nitriles”、“fidarestat”、“inaclotide”、“silatrane” などがあげられる。

6. まとめ

本稿ではテキストからの化合物名抽出において、サブワードの埋め込みベクトルをモデルに加えることで抽出精度が上がることを示した。

今回は語彙数が 2,000、4,000、16,000 に限定し実験を行ったが、それぞれで抽出の精度が異なっていた。サブワードの語彙数を決定することは、このモデルの重要な要素の一つでもあるため、今後は語彙数と抽出精度の関係性について深く調べる必要がある。

また、今回の実験では語彙数が 4,000 と 16,000 の二つのサブワード系列を入力した場合に最も良い精度となった。しかし、そのモデルに対し、語彙数が 2,000 のサブワード系列を

加えた場合の抽出精度は芳しくなかった。これは、単純にサブワード系列を加えていくのではなく、どのサブワード系列を使用するか判断する必要があることを示している。

今後は、加えるサブワード系列の語彙数と、どの語彙数を加えると良いスコアを得るかについてもう少し、研究していくことが必要である。

参考文献

[1] Krallinger et al. (2015) The CHEMDNER corpus of chemicals and drugs and its annotation principles. J Cheminform. 2015 Jan 19;7(Suppl 1 Text mining for chemistry and the CHEMDNER track):S2. doi: 10.1186/1758-2946-7-S1-S2. eCollection 2015.

[2] Jie Yang et al. (2018) Design Challenges and Misconceptions in Neural Sequence Labeling. 2018 13 Aug. CoRR. abs/1806.04470

[3] Ling Luo et al. (2018). An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics, 2018 Apr 15;34(8):1381-1388.

[4] Alan Akbik et al. (2018) Contextual String Embeddings for Sequence Labeling. 2018 Aug. Proceedings of the 27th International Conference on Computational Linguistics, p.16381649

[5] Jeffrey Pennington et al. (2014) GloVe: Global Vectors for Word Representation. Empirical Methods in Natural Language Processing (EMNLP). p1532–1543. https://github.com/stanfordnlp/GloVe

[6] Taku Kudo et al. (2018) SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. 2018 Aug. CoRR. abs/1808.06226. https://github.com/google/sentencepiece

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

文書からの化合物名抽出のためのサブワード有効性調査

Using Subword Sequence BiLSTM-CRF Model for Compound Name Extraction

浦澤合 *1　関根裕人 *1　乾孝司 *1　岩倉友哉 *2
Go Urasawa　Hiroto Sekine　Takashi Inui　Tomoya Iwakura

*1筑波大学大学院　/理研 AIP-富士通連携センター
University of Tsukuba/RIKEN AIP-FUJITSU Collaboration Center

*2富士通研究所/理研 AIP-富士通連携センター
Fujitsu Laboratories/RIKEN AIP-FUJITSU Collaboration Center

In this paper, we investigate of using subword sequences for compound name extraction problem. Five variety of subword sequence generators (SYMBOL, SP, BPE, BPE-DICT, and BPE-PMI) were used in the investigation. Last two of these, BPE-DICT and BPE-PMI, are originally proposed in this work. BPE-DICT is a variation of BPE which has a dictionary-based restriction. BPE-PMI introduces the PMI measure instead of word frequency count. The experimental results showed that subword sequence information improved the extraction performance. The F-measure value of BPE-DICT is 86.74 which is best score in all conditions of our experiments.

1. はじめに

人間の言葉で書かれた文書から自動的に構造化データを抽出するタスクを「情報抽出」という．抽出するデータの種類に応じてデータの特徴，文書形式などが異なる．したがって抽出データの分野により情報抽出をおこなうのに最適な抽出手法や考え方が異なるので抽出データに合わせた手法を考える必要がある．本研究では化学化合物に関する情報抽出について考える．

化学分野の研究ではさまざまな場面で化合物データベースが利用され，日本化学物質辞書 [5] や PubChem[4]，ChEMBL[1] など各種データベースが提供されている．現在，これらの化合物データベースは論文や特許を人手で読み解くことで作成されている．しかし化学化合物に関する論文，特許は数え切れないほどの数が存在するのでそれらから必要なデータを人手で抽出することは非常にコストの高い作業である．そこで解析技術と組み合わせたデータベース作成支援が求められている．

化合物抽出用のソフトウェアが存在するが，精度の点で実用レベルに至っているとは言い難い．化学化合物の抽出が実用レベルでないのは，化学化合物の命名規則である IUPAC 命名規則 [2] があるにも関わらず化合物の多様な表記方法があることが理由として挙げられ，略称，通称，化学式など多くの表記を持つ．例として化合物「フェニルアラニン」は「Phe-OH」,「L-Phe-OH」,「L-Phenylalanine」,「(S)-2-Amino-3-phenylpropionic acid」などで表現されるが，これで全ての表記ではない．また化学の分野において新しい化合物が頻繁に報告されることも理由として挙げられる．他には複合語の存在である．複合語とはある化合物を部分的に含む化合物のことである，これらにより化合物の正確な抽出が困難となる．

上記で述べたように，化学化合物は未知語が発生しやすい分野であるため，未知語に対応する処理が重要となる．サブワードは単語と文字両者の中間的な特性を持つため単語情報を保持しつつ，未知語に対応できると考えられる．そのため本研究は様々な分割方法でサブワードを獲得し，どのようなサブワード系列が化学化合物抽出において効果が見られるか検討する．

連絡先: 浦澤合，筑波大学大学院システム情報工学研究科，
g.u@mibel.ca.tsukuba.ac.jp

本稿の構成として，第 2 章で関連研究について述べる．第 3 章ではサブワード獲得方法について述べる．第 4 章では評価実験について述べる．最後に，第 5 章では予備調査を踏まえた考察と今後の指針を述べる．

2. 関連研究

化学分野の論文や特許から化合物を抽出する研究 [8][9][6] は盛んに行われている．機械学習を利用した手法 [8] が高精度の結果を残すものとして以前から知られていたが，近年ではニューラルネットワークを利用した手法 [9] が多く提案され化合物抽出において素晴らしい結果をもたらしている．

機械学習を利用した Lu ら [8] は文字や単語から獲得できる情報を CRF の素性として利用した．また単語のクラスタリングを素性として利用することで精度や再現性を高めようとした．次に Ling ら [9] は化合物抽出のニューラルネットワークを利用した手法によく見られる BiLSTM に attention 機能を追加した手法を提案し，CHEMDNER task[6] では現在もっとも精度の高い手法となっている．以上のように単語および文字の情報を利用した手法は数多く提案されているが，本研究のようにサブワードを利用した手法は提案されていない．

3. サブワード

3.1 固有表現抽出におけるサブワード

化合物名抽出を行う際には系列ラベリング問題として定式化することが一般的である．系列ラベリング問題として考えた場合に単語単位の系列を仮定すると，未知語が発生しやすくなり，また，抽出したい化合物と処理上の単位である単語との間で境界が一致しない問題を引き起こす．この問題への対策として，文字単位系列を用いることが考えられるが，この場合，系列長が長くなり計算量が増加する．また，単語がもっていた意味情報を利用することができないといった新たな問題が発生する．単語の使用と文字の使用はそれぞれに利点と欠点があり，両者はトレードオフの関係にあると言える．サブワードはこの両者の中間的な特性をもっており，サブワードを考慮した系列を仮定することで，単語と文字の両者の欠点を補うことができ

Algorithm 1 BPE

1: $DICT \Leftarrow$ 辞書データ
2: $VOCAB \Leftarrow$ 語彙 (初期は空)
3: テキストを文字に分割する
4: **while** $VOCAB$ が指定語彙数に達するまで **do**
5: 隣り合う全ての文字トークンのペアに対して，それらを結合して新たな語彙候補を作成する．($VOCAB$ に存在するものは一文字として扱う)
6: 語彙候補の中で，出現頻度の最も高い候補を $VOCAB$ に追加する．
7: **end while**

Algorithm 2 BPE-DICT

1: $DICT \Leftarrow$ 辞書データ
2: $VOCAB \Leftarrow$ 語彙 (初期は空)
3: テキストを文字に分割する
4: **while** $VOCAB$ が指定語彙数に達するまで **do**
5: 隣り合う全ての文字トークンのペアに対して，それらを結合して新たな語彙候補を作成する．($VOCAB$ に存在するものは一文字として扱う)
6: 結合前の左右の要素がどちらも $DICT$ に登録されていない語彙候補の中で，出現頻度の最も高い候補を $VOCAB$ に追加する．
7: **end while**

ると考えられる．サブワードとは，ある単語の部分文字列のことである．例として「magnesium」という単語の場合,「ma」,「mag」,「magn」,「si」,「sium」などがサブワードになる．

3.2 サブワード獲得方法

本研究では単語から得られるサブワードとして以下 5 種類のサブワードを試みた．このうち，BPE-DICT と BPE-PMI は本研究で提案するサブワード獲得方法である．

- 記号などが存在する際にその記号で単語を分割するもの (SYMBOL),
- SentencePiece(SP),
- Byte Pair Encoding(BPE),
- 辞書制約付き Byte Pair Encoding(BPE-DICT),
- PMI による Byte Pair Encoding(BPE-PMI).

3.2.1 SYMBOL

SYMBOL は単語中に記号などが存在する際に，その記号で単語を分割するサブワード分割方法である．一般的な単語は記号を単語中に含まないので，単語は分割されず，単語そのままであることが多い．単語が組み合わせられた複合語や，長い単語には記号が含まれることが多く，分割される．

3.2.2 SP

SentencePiece[3][7] を実行することでサブワード系列を獲得する．

3.2.3 BPE

Byte Pair Encoding は Sennrich ら [10] が提案したサブワード分割方法である．BPE は原文をすべて文字に分割し，1 文字 1 語彙から始まる．隣り合う文字のペアに対して，それらを連結して新たな語彙の候補とする．この際すでに語彙に含まれているものは 1 文字として扱う．連結した際に最も出現頻度が高くなるサブワードを選び語彙に追加する．この手続きを決められた語彙数に達するまで繰り返し語彙結合ルールを学習することでサブワード分割を行う．

3.2.4 BPE-DICT

BPE-DICT は辞書制約付きの BPE である．基本的な手続きは通常の BPE と同じであるが，連結する前の左右の要素どちらも化学化合物辞書に存在しない語彙候補の中で出現頻度が最も高いものを語彙に追加する．また今回，化学化合物辞書として利用したのは PubChem データベース [4] で約 3 億個の化合物を含んでいる．

3.2.5 BPE-PMI

通常の BPE は出現頻度が高いサブワードを新たな語彙として追加するが，BPE-PMI は出現頻度ではなく Pointwise Mutual Information(PMI) が高いサブワードを新たに語彙に追加する．学習データを参照することで各サブワードについて，化合物の構成要素 (クラス 1)，構成要素でない (クラス 0) を割り当て，各サブワードとクラス 1 間の PMI を求め，出現頻度に置き換えて BPE を行う．式 (1) に PMI の定義式を示す．ここで，$P(SW)$ はあるサブワードが出現する確率，$P(C=1)$ はある要素が化合物の構成要素である確率，$P(SW, C=1)$ はあるサブワードが出現した際にそれが化合物の構成要素である確率である．

$$PMI(SW, C=1) = \log 2 \frac{P(SW, C=1)}{P(SW)P(C=1)} \quad (1)$$

4. 評価実験

4.1 実験設定

前節で述べた手法によって得たそれぞれのサブワード系列が化学化合物抽出にどの程度有効であるか観察した．データセットは CHEMDNER tsak における学習データ 3,500, 開発データ 3,500, 評価データ 3,000 を利用した．これは PubMed から化合物について書かれた論文の abstract を 10,000 件集め，人手でアノテーションをされたものである．合計で 84,355 の化合物エンティティが存在し，それらの重複を省くと 19,806 となる．化学化合物の固有表現抽出モデルとして Bidirectional LSTM-CRF[9] を使用し，これは単語と文字の LSTM を持つ．本研究では固有表現抽出モデルの単語 LSTM をサブワードに

表 1: サブワード別出力例：学習データ内に存在する化合物

method ＼化学化合物	docosahexaenoic acid	nitric oxide	glutathione	superoxide
SYMBOL	docosahexaenoic acid	nitric oxide	glutathione	superoxide
SP:chem6k	docosahexaeno ic acid	nitr ic oxide	glutathione	super oxide
BPE:32k	docosahexaenoic acid	nitric oxide	glutathione	superoxide
BPE-DICT:32k	docosahexaenoic acid	nitric oxide	glutathione	su p eroxide
BPE-PMI:32k	docosahexaenoic acid	nitric oxide	glutathione	superoxide

表 2: サブワード別出力例：学習データ内に存在しない化合物

method ＼化学化合物	isocorilagin	diasesartemin	tetrahydropalmatine	ritanserin	polyphosphoinositides
SYMBOL	isocorilagin	diasesartemin	tetrahydropalmatine	ritanserin	polyphosphoinositides
SP:chem6k	isoc or il ag in	di a s es ar te m in	tetrahydro p al mat ine	rit an serin	poly phosphoinositide s
BPE:32k	is ocor il agin	di as es artem in	tetrahydro palmatine	rit anserin	poly phospho inosi tides
BPE-DICT:32k	isoc or il agin	di as es artem in	tetrahydro pal matine	rit an serin	p oly p hos p hoinosi tides
BPE-PMI:32k	isoc ori lag in	di as esar te min	tetrahydro palmatine	rit ans erin	polyphospho inositi des

表 3: 利用したモデルのパラメータ

epoch	200
batch size	100
単語分散表現	50
文字分散表現	30
単語 LSTM の隠れ層	100
文字 LSTM の隠れ層	50
initial rate	0.015
dropout	0.5

置き換えて用いる．また今回使用したモデルのパラメータを表 3 に示す．単語分散表現，文字分散表現はそれぞれ 50 次元，30 次元とし，LSTM の隠れ層では 100 次元，50 次元とした．表中の「単語」は実際にはサブワードである．

サブワード獲得方法別の設定を述べる．SentencePiece の学習には上記と同じ学習データを利用した．学習データの全テキストを語彙数 32,000，ユニグラムで学習させたもの (SP:32k) と学習データのタグづけされた化学化合物部分のみを語彙数 6,000，ユニグラムで学習させたもの (SP:chem6k) がある．次に BPE,BPE-DICT,BPE-PMI について説明する．これら 3 つも同様に先と同じ学習データを利用したが，BPE,BPE-DICT は学習データのテキスト部分のみを語彙獲得に利用し，BPE-PMI は学習データのテキスト部分とタグ部分を語彙獲得に利用した．また 3 つそれぞれに対して 8,000，16,000，32,000 の語彙数で学習させた．また BPE-DICT では辞書引きを行う手続きに仮候補の文字列が 3 文字以上という制限を加えたもの (BPE-DICT-char3) と制限なしのもの (BPE-DICT) がある．この制限は制限なし BPE-DICT が獲得したサブワードを観察した際に，多くの 1 文字サブワードが残った．その結果に対して調整を行う目的で取り入れた．

4.2 実験結果と考察

表 4 に実験結果を示す．性能を F-measure の値で比較すると，最も良い性能であるのは語彙数 32,000 で制限なしの BPE-DICT である．これの F-measure は 86.74 であり，ベースラインである単語区切りの 86.32 に 0.4 上回っている．これは通常の BPE と比較しても性能が良いことから辞書制約付きがある BPE は化合物抽出において効果があると言える．

またシンプルなサブワード分割方法である SYMBOL もベースラインの F-measure を上回っており，これはサブワードが単語よりも化合物抽出で良い影響を持っていると言える．BPE-PMI の F-measure は通常の BPE と比べて低い．したがって各サブワードと化合物の構成要素クラス間との PMI をもとに BPE を行うよりも出現頻度をもとにした通常の BPE の方が今回の実験では良いサブワード分割ができていると言える．

表 4: サブワード別実験結果

method	Precision	Recall	F-measure
単語	86.62	86.03	86.32
SYMBOL	85.10	88.01	86.53
SP:32k	87.56	85.39	86.46
SP:chem6k	78.90	69.53	73.95
BPE:8k	87.12	84.98	86.04
BPE:16k	87.47	85.99	86.72
BPE:32k	84.38	79.38	81.80
BPE-DICT:8k	87.74	84.32	86.00
BPE-DICT:16k	87.37	85.70	86.53
BPE-DICT:32k	87.14	86.33	86.74
BPE-DICT-char3:8k	87.17	84.33	85.73
BPE-DICT-char3:16k	86.40	86.39	86.39
BPE-DICT-char3:32k	88.10	85.10	86.62
BPE-PMI:8k	85.51	82.45	83.95
BPE-PMI:16k	86.86	82.76	84.76
BPE-PMI:32k	78.69	71.03	74.66

SP:chem6k と BPE-PMI:32k の性能が他と比較すると極端に低い．SP:chem6k と BPE-PMI:32k ともにサブワード分割の特徴として化学化合物が分割されることが少なく，残りやすいことが挙げられる．これは学習をともに学習データの化学化合物部分に集中して行うからと考えられる．化学化合物とそれ以外の違いを他のサブワード分割方法では区別できているとも言える．

表 1 と表 2 に各サブワード分割方法で得られた化学化合物のサブワード出力例をいくつか示す．表 1 の化学化合物は学習データに化合物部分としてタグ付けされているのでほとんどのサブワード分割方法が分割を行わず単語を維持している．反対に，表 2 の化学化合物は学習データに化合物部分としてタグ付けされていないため多くのサブワード分割方法で分割が行われる．また単語や記号でサブワード分割を行うシンプルな分割方法である SYMBOL では表 2 の化合物すべてが抽出するこ

とができなかった．しかし実験結果で最も F-measure が高いサブワード分割である BPE-DICT:32k では表 2 の化合物を抽出することができた．したがって，学習データに化合物部分としてタグづけされていない化合物でもサブワードを利用することで抽出することを可能とし，DICT-BPE:32k は良いサブワード分割を行なっていると言える．

5. おわりに

今回，様々なサブワードを Bidirectional LSTM の入力系列として利用し化学化合物抽出を行い，各サブワード獲得方法別の結果を観察し，サブワードが単語よりも化学化合物抽出において効果があることを示した．

結果としては BPE-DICT が最も高い性能を残したが，各サブワード獲得方法すべてに同様のパラメータを適用したので BPE-DICT が化学化合物抽出に最も適しているとは一概にも言えない．さらに今回語彙数として 8,000，16,000，32,000 を使い実験を行なった，それに対して実験結果や結果の傾向が異なるものとなったが，これはサブワードを利用する上で語彙数を決定することは化学号物抽出の性能に大きな影響があると言える．今後は各サブワード獲得方法に適したパラメータの調査やサブワード分割方法とそれに適した語彙数の決定について行う必要がある．また以上のことから得られる化学化合物抽出に適したサブワードをサブワードのみの入力系列ではなく，単語系列などと組み合わせた実験などを行なっていきたい．

参考文献

[1] ChEMBL. https://www.ebi.ac.uk/chembl/.

[2] Color books-iupac international union of pure and applied chemistry. https://iupac.org/what-we-do/books/color-books/.

[3] Github-google/sentencepiece: Unsupervised text tokenizer for neural network-based text generation. https://github.com/google/sentencepiece.

[4] The pubchem project. https://pubchem.ncbi.nlm.nih.gov/.

[5] 日本化学物質辞書 web—j-global 科学技術総合リンクセンター. https://jglobal.jst.go.jp/info/nikkaji.

[6] Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. Chemdner: The drugs and chemical names extraction challenge. *Journal of cheminformatics*, Vol. 7, No. 1, p. S1, 2015.

[7] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

[8] Yanan Lu, Donghong Ji, Xiaoyuan Yao, Xiaomei Wei, and Xiaohui Liang. Chemdner system with mixed conditional random fields and multi-scale word clustering. *Journal of cheminformatics*, Vol. 7, No. S1, p. S4, 2015.

[9] Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics*, Vol. 34, No. 8, pp. 1381–1388, 2017.

[10] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

運転免許試験で使用される語彙と省略語句の分析

Analysis of vocabulary and ommitted words in car license tests

的場 成紀 *1 Seiki Matoba　古賀 雅樹 *1 Masaki Koga　大塚 基広 *1 Motohiro Otsuka　小林 一郎 *2 Ichirou Kobayashi　平 博順 *1 Hirotoshi Taira

*1大阪工業大学大学院 情報科学研究科
Faculty of Information Science and Technology, Osaka Institute of Technology

*2お茶の水女子大学大学院 人間文化創成科学研究科
Graduate School of Humanities and Sciences, Ochanomizu University

We develop a solver for Japanese car license tests. The test consists of about a hundred of true/false questions about traffic rules, driving manners, architectures of cars and the laws of physics related to cars. While the passing score is 90%, The best score in the previous approaches is about 65%. The approach is based on the sentence similarity between the test sentence and most similar sentence with the gold-standard answer in the database in the solver. Toward the system to pass the test, we analyzed the vocabulary and writing styles of the tests. The results of the analysis showed that the vocabulary is relatively small, which is about 300 words for 100 problems, and the sentences contain a lot of zero pronouns and they cause the low accuracy of the solver. Furthermore, we tried to resolve the antecedents using a previous anaphora resolution system. The results showed that the system cannot resolve the anaphora in the tests, because each problem consists of only one sentence and the clue to resolve the pronoun is very few, and they are more difficult to resolve than ones in standard articles. The analysis has revealed that high-performance systems require the anaphora resolution which is more based on domain specific knowledge.

1. はじめに

自動車免許試験に対して自動解答するシステムの検討を行っている [平 14]．本研究は自動車免許試験を自動解答するソルバを開発するために免許試験のコーパスの分析を行う．これまでの研究では，問題文とソルバが持っている問題データベース中の問題文との間の単語類似度を利用して正誤判定を行う手法が提案されている [杉村 13]．この手法では，図を使用しない自然文のみからなる模擬問題に対して，約 6 割の正解率が得られている．また，従来の自動車免許試験についての問題分析では，問題のトピック，問題の言い回し，問題を解くために必要な技術や知識に関する検討が行われている [平 15]．本研究では，免許試験問題に対して自動解答する上で必要な語彙について 5W1H の観点から分析するとともに，問題文中で頻出する省略語句について分析を行った．

2. 普通自動車免許学科試験問題の概要

普通自動車免許を取得するためには，実際に自動車を運転し，運転技術について評価する実技試験と交通規則やマナー等，運転する上で必要となる知識を問う学科試験の 2 つに合格する必要がある．後者の学科試験は，制限時間 50 分の筆記試験である．出題形式は正誤判定問題であり，交通規則やマナー，運転知識などについて述べられた問題文に対し，そこで述べられていることが正しいか誤っているかを判定する問題である．問題は 95 問出題され，先頭 90 問は，問題文各 1 文を読み正誤判定を行う問題である．自然文の問題文だけを読み解答する問題が多いが，交通標識などのイラストの絵を参照して解答する問題も存在している．それに対し，末尾の 5 問は，1 問につき 1 枚の運転席から見た外の様子などのイラストが示され，3 つの枝問の正誤判定問題を解く問題である．先頭 90 問は各 1 点，末尾 5 問は完答で各 2 点で，合計 100 点満点である．学科試験の合格基準は 90 点以上である．

連絡先: 的場成紀, 大阪工業大学情報科学部, 〒 573-0171 大阪府枚方市北山 1-79-1, e1b15097@st.oit.ac.jp

3. 学科試験問題で使用されている語彙の分析

3.1 問題文中の語彙数

今回，分析の対象とした学科試験の模擬問題は「試験によく出る普通免許 1000 題」（倉 宣昭著，高橋書店）の第 10 回の問題の問題文 100 文とした．まず，問題文中に含まれる内容語の語彙数について調査したところ，319 語であった．ただし，「原動機付自転車」などの語は「原動機」「付」「自転車」などと 3 つには分けず 1 語として扱った．

3.2 問題のタイプ

3.2.1 自動解答処理の観点から見た問題分類

試験問題には異なったタイプのものが存在する．そのため，解答する際にはどこに注目するべきなのかを理解する必要がある．まず，試験問題は次の通り大きく 3 種に分類する事が出来ると考えた．まず 1 つ目はイラストを見て解答する問題である（以下「イラスト問題」と略す）．分析対象の 100 問中にはイラスト問題は 13 問存在した．

2 つ目は問題文で示された物理現象について，法則に基づいて計算が必要な問題である．この問題では，問題文が問われている法則の種類，計算に必要な情報を特定した上で，計算を行った上で正誤判定を行う必要がある．これは 100 問中，2 問存在した．

3 つ目はそれ以外の問題である．この問題は，イラストはなく，問題の自然文で記されている交通ルールやマナーなどについての記述が正しいかどうかを判定する問題である．この問題は，100 問中，85 問存在した．

3.2.2 詳細な分類

以下の異なる 3 つの観点で，問題をより詳細に分類した．

1. 5W1H による分類
2. 問われ方による分類
3. 問題の内容

3.2.3 5W1H による分析

問題文中に出現する内容語の語彙について 5W1H の観点から分析を行った．ここで 5W1H は，時間 (when)，場所 (where)，ガ格 (who)，ヲ格 (what), 述語（how) とした．

(1) when: 季節や天候，時間帯，災害時なども含めた時期や時間

(2) where: 場所

(3) who: 文中のガ格に相当する内容語

(4) what: 文中のヲ格に相当する内容語

5W1H の観点から内容語を分類したものを表 1 に示す．

まず,「時間」の中で最も多く出現した単語は「夜間」であった．これは，視界や前照灯に関する問題において,「夜間」という単語が多く用いられているためである．

「場所」に関しては単語「道路」が最も多く出現していた．また,「道路」という単語は，単独で使用されることは少なく,「一方通行の」や「車両通行帯のある」などの修飾語句を伴って使用されることが多かった．

「ガ格」については,「運転手」「自動車」などが多く出現していたが,「ブレーキ」「後車輪」などの自動車の部品に関する単語も多く見られた．また，ガ格である単語が直接的な問題の対象物ではない問題も多く見られた．例えば，

- ファン・ベルトの中央を指で押してみたら、50 ミリメートルぐらいの緩みがあった

の問題では，述語「あった」に対するガ格は「緩み」であるが，問題で問われている知識は,「ファン・ベルトの緩みの大きさ」についての知識であり，問題が問うている対象の解析を難しくしていることが分かった．

3.2.4 問題の問われ方による分析

次に，問題の問われ方について分析を行った．自動車免許試験の正誤判定問題の言い回しは，ある種独特のものであり,「〜は〜である」といった「事実について述べている文」と「〜は〜した」といった運転手や車などの「行動について述べている文」の 2 種類が存在し，それらの割合について分析を行った．

分析の結果，表 2 で示す通り,「事実について述べている文」の方が多く，出現数は 84 であった．それに対して,「行動について述べている文」の出現数は 16 であった．さらに，それらの文についての文末表現について分析したところ,「行動について述べている文」に関しては最後が「〜した」で終わることが多く,「事実について述べている文」には「〜である」「〜できる」「〜になる」「〜よい」など，表現に多様性が見られた．

3.2.5 問題の内容による分析

問題の内容による分析を行う．

大きく分ける種類として 4 種類あり，それぞれ「運転」「準備」「知識」「心がけ」である．運転は運転に関することに対して問われている問題である．例えば「追い越し」などである．

表 1: 5W1H の観点から分類した内容語と頻度

5W1H	単語 (頻度)
when (時間)	夜間 (6), 雨 (3) 冬 (1), 雪 (1) 災害 (1), 光化学スモッグ (1)
where (場所)	道路 (12), 交差点 (6) 高速道路/高速/高速自動車国道 (4), 坂 (4) 踏切 (3), 路側帯 (2), 踏切 (2) 曲がり角 (1), 市街地 (1), トンネル (1)
who (ガ格)	運転手 (4), 速度 (3) 自動車/車/普通自動車 (3), 者 (3) エンジン・オイル (2), 空気圧 (2) 原動機付き自転車/原動機付自転車 (2), 追い越し (1) 警察官 (1), 交通 (1) 光化学スモッグ (1), 交通公害 (1) 後車輪 (1), 視線 (1) 灯火 (1), 燃料 (1) (オイルの) 循環状態 (1), (燃料の) 消費量 (1) タイヤ・チェーン (1), トンネル (1) (排気の) 色 (1), ハンド・ブレーキ (1) 普通二輪車 (1), ブレーキ (1) 前車輪 (1), (タイヤの) 溝 (1) ミニカー (1), 路面電車 (1)
what (ヲ格)	自動車 (8), オートマチック車 (8), 車 (8) ブレーキ (3), 急ブレーキ (3) ペダル (3), ブレーキ・ペダル (3) 原動機付き自転車/原動機付自転車 (2) 人 (2), 2 人 (2), 大型特殊自動車 (1) 違反 (1), 後車輪 (1) 燃料 (1), エンジン (1) ファン・ベルト (1), 速度 (1) 幼児 (1), 前照灯 (1) 急ハンドル (1), 視力 (1) チェンジ・レバー (1), 進路 (1) 通行帯 (1), 路側帯 (1) 危険 (1), 全引きしろ (1) ハイドロ・プレーニング現象 (1) 車間距離 (1) 速度超過 (1), 余地 (1) 積載超過 (1), 間隔 (1) 場所 (1), マイクロバス (1) 許可 (1), 車線 (1) 故障車 (1), 番号標 (1) 荷物 (1)

表 2: 問題の問われ方による分類の結果

問題の問われ方	頻度
事実について述べている文	84
行動について述べている文	16

表 3: 問題の内容による分類の結果

問題の内容	頻度
運転の技術	77
準備	17
知識	4
心がけ	2

準備は運転をする前に行う前の段階である．知識は運転に関する現象である．心がけは道徳的な話が含まれている．

表 3 にこれらの分類の頻度数を示す．一番多く出題されているのは「運転の技術」についてのものである．

次に多く出題されていたのは「準備」であった.「準備」に関する問題は，主に車の部品に関することや免許の種類に関する内容である．

「知識」に関する問題は，運転中におこる現象に関する内容や数式に分類される問題である．

「心がけ」は出題される問題が一番少なかった．

3.3 問題を解く上での必要な知識の分析

問題を解くために必要な知識を考察する．先の問題の分類を踏まえると

1. 運転に関する知識
2. 常識的な知識

の 2 種類が必要であると考えられる．

運転に関する知識とは「交通の方法に関する教則（昭和 53 年 10 月 30 日　国家公安委員会告示第 3 号)」に掲載されている知識について問われている．例えば

- 車の速度が 2 倍になると、制動距離はおおよそ 4 倍になる。

上記のような問題の場合,「制動距離」という単語の意味を知らなければならない．加えて，制動距離の算出の仕方も知っている必要がある．

「常識的な知識」とは運転に関する知識とは違い，明記されていない情報を想像するために必要な知識である．例を挙げると

- オートマチック車で坂を下るときは、チェンジ・レバーを 2 か 1 に入れ、エンジン・ブレーキを活用する。

上記の問題では動作主格となる単語が文中にない．しかし，人間の受験者は，問題文を見たときに「動作主格が運転手である」と容易に想像することができる．これは人が今までに培ってきた経験から推測できるためである．

3.4 問題文の省略解析

運転免許問題には省略語句が多く含まれる．問題文に対して，省略解析を行って分析を試みた．省略解析器として KNP を使用した．

表 4 に KNP による解析結果を示す．100 問の問題文中で，省略格指定が 17 個，省略解析対象指示詞が 11 個，合計 28 個の省略があった．省略格指定では主に「見える」の単語が対象となった．例えば

- 暗いトンネルから明るい場所へ出たときは、視力が急激に低下して、見えなくなることがある。

表 4: 省略解析の結果

ゼロ照応解析の結果	頻度
省略格指定	17
省略解析対象指示詞	11
合計	28

上記のような問題であった場合,「見えなくなる」の部分が省略格指定となる．また,「不特定人が」,「トンネルから」,「場所」が直接係り受け解析の格の対象となる．不特定人とは問題文の中に明記はされていないが主格の対象となるのは人であるときの対象となる．また,「とき」という単語が明記されていないが「見えになくなる」に係り受け解析の結果となっており,「見える」が何に対して指示対象としているのかを表す単語は「視力」となっている．

上記以外にも「〜という」「〜といえる」の単語が省略格指定の対象となった．例えば

- 運転者は、「酒を飲んだら運転しない」「乗るなら飲まない」という習慣を身につけることが大切である．
- 交通法令を守らなくても、臨機応変に運転して交通事故を起こさなければ、「安全な運転」といえる．

である.「〜という」の場合，主に主格となる部分は不特定人となった．また,「〜といえる」の場合は主格が不特定人となるが，上記の例では「運転と」が省略されていると特定された．

上記の KNP で省略解析を行った結果と予測していた結果と異なっていた．その原因の一つとして，免許問題が 1 文で構成されることが挙げられる．本来，KNP で省略解析を行う場合は 2 文以上あることを想定している．しかし，今回のように問題文が 1 文しかないために予想とは違った結果となったと考えられる．

4. おわりに

本稿では問題文の分析を 5W1H の観点からの使用語彙と省略語句について分析を行った．5W1H の観点で使用語彙を見た場合,「時間」は「夜間」,「場所」は「道路」,「ガ格」は「運転手」,「ヲ格」は「自動車」が一番多く頻出した．さらに省略語句について省略解析器を用いて問題文の解析を行い分析を行った．今後は，上記の結果を踏まえて自動解答システムを開発する予定である．

謝辞

本研究は JSPS 科研費 18K11452 の助成を受けたものである．

参考文献

[杉村 13] 杉村 皓太, 佐々木 裕 : 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 発表論文集, pp. 790–793 (2013)

[平 14] 平 博順, 田中 貴秋, 永田 昌明 : 自動車運転免許試験 RTE コーパスの構築, 第 28 回人工知能学会全国大会予稿集, 3I4-5 (2014)

[平 15] 平 博順 : 自動車免許試験自動解答に向けた問題分析, 第 29 回人工知能学会全国大会予稿集, 1K2-2 (2015)