専門家と非専門家によるアノテーション検証の割当の自動化

守山慧 [1]　中山功太 [2,4]　馬場雪乃 [3,4]
[1] 筑波大学情報学群情報科学類　[2] 筑波大学大学院理工情報生命学術院
[3] 東京大学大学院総合文化研究科 [4] 理化学研究所 AIP
s2113576@s.tsukuba.ac.jp　kouta.nakayama@riken.jp
yukino-baba@g.ecc.u-tokyo.ac.jp

概要

自然言語理解を実現するために、知識ベースの構築が必要である。人手による知識ベースの構築にはコストがかかるため、機械学習モデルを用いて自動化することができる。だが、機械学習モデルにより作成された知識ベースは正しいとは限らないので検証する必要がある。

これを達成するために、本研究では、適切な問い合わせ先として機械、非専門家、専門家のいずれかを自動選択する手法を提案する。実データを用いた実験で、提案手法により問い合わせ先を選ぶ場合と、ランダムに問い合わせる場合とで比較したところ、提案手法の方がより精度良く検証が実現できることを確認した。

1　はじめに

自然言語理解を実現するためには、言語的及び意味的な知識ベースの構築が必要になる。このような知識ベース構築プロジェクトの１つとして、森羅プロジェクト [1] がある。[1] では Wikipedia の記事を用いて知識ベースの構築を目指している。Wikipedia の記事から固有表現を抽出することで計算機が利用可能な形式での Wikipedia の構造化を進めている。

しかし、Wikipedia のようなテキストの更新が頻繁に行われ、数が膨大なテキストから知識ベースを人手で構築することは困難である。この問題の解決策として、機械学習モデルを用いて、テキスト中に含まれる固有表現を抽出し、知識ベースの構築を自動化することでコストを下げることが挙げられる。

機械学習モデルによる自動化を行った際の課題として、抽出された固有表現が正しいとは限らないという点が挙げられる。そのため、構築された知識ベースが実用的ではなくなってしまう。これを防ぐために、機械学習モデルが抽出した固有表現を検証し、間違って抽出された固有表現が混ざることを防ぐ必要がある。だが、人手による検証は数が膨大であるため困難である。よって、人手による検証をする必要のあるデータと、検証する必要のないデータを選択することで検証にかかるコストを下げることが必要である。

また、人手による検証をする際の課題として、タスクに対する知識の量は人によって異なる点が挙げられる。タスクについて専門的な知識を持つ人を専門家、そうではない人を非専門家とすると、専門家はタスクに対して深い知識を持っているため、正確な解答をすることができる。一方で、非専門家は難しいタスクに対して正確ではない解答をする。また、解答を依頼する際のコストは、専門家の方が多く、非専門家の方が少なくい。そのため、専門的な知識を必要としないで判定が可能なデータについては非専門家、必要とするデータについては専門家に検証を依頼することが理想的である。

そこで本研究では、固有表現の抽出結果の検証方法の選択を行うモデルを学習させる手法を提案した。検証方法として、機械、非専門家であるクラウドワーカー、専門家であるアノテーターの 3 種類を対象とした。この損失関数で学習させた機械学習モデルと、ランダムにデータの検証方法を決定した場合を比較して正解率、F1 値、再現率、精度の比較をした。学習させた結果、学習させたモデルの方が正解率が高く、また割当先をうまくコントロールすることで F1 値もランダムにデータの検証方法を決めるよりも良くなっていることがわかった。

2　関連研究

人間が解答するデータと機械学習モデルによる予測を行うデータの割当を行う手法として [2, 3, 4, 5] がある。これらの手法は主に２つのアプローチがある。予測を行う前にデータを分割する手法 [2, 3, 5]

と、予測とタスクの依頼を同時に行う手法 [4] がある。Raghu ら [2] の手法では、人間の予測に対する損失関数と機械学習モデルの予測に対する損失関数をそれぞれ定義し、この損失が最小になるようにデータを分割することを提案している。Nastaran ら [3] は、人間に依頼するかどうか決定するモデルと分類モデルを同時に学習させる手法を提案している。これらのモデルを学習させる前に、訓練データの中から分類モデルの学習に使うためのデータを分割し、学習を行った後、訓練データを用いて人間に依頼するかどうかを決定するモデルの学習を行う。Wilder ら [5] の手法では、あるデータについて機械学習モデルを用いて予測を行うか、人間に問い合わせて答えを得るかを選択する問い合わせモデルを学習させるための損失関数を提案している。

既存研究では、人間に問い合わせを行う際に専門的であるかどうかは考慮せず、問い合わせを行う人は専門的な知識を持っているという仮定をしている。本研究では、人間への問い合わせ先として、専門的な知識を持つ人と持たない人の 2 つの問い合わせ先を用意した点が既存研究と異なる。

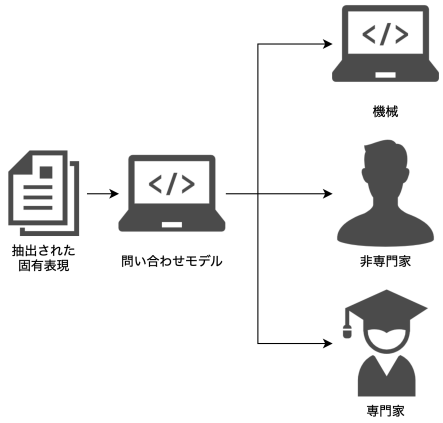

図 1　問題設定の概略図

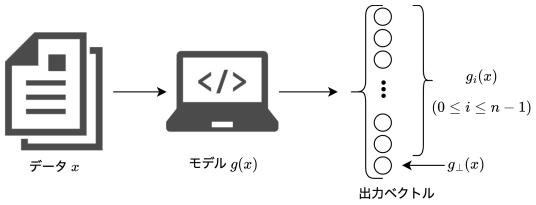

図 2　既存手法 [4] のモデルと出力ベクトルの概略図

3　問題設定

Wikipedia のページから抽出された固有表現の検証タスクを、機械、非専門家であるクラウドワーカー、専門家であるアノテーターのいずれかに問い合わせる問い合わせモデルの学習を目指す。問題設定の概略図を図 1 に示す。例えば、Wikipadia のページから固有表現抽出を行った結果「所在地」クラスの固有表現として「モエリス湖」が抽出されたとする。この固有表現に対して、機械、クラウドワーカー、アノテーターのいずれかが検証を行う。検証結果は $\{0, 1\}$ であり、1 が正解、0 が誤りである。

それぞれの検証方法にはメリット、デメリットがある。機械における検証はコストがかからないが、判定結果に信頼性が低い。非専門家のクラウドワーカーによる検証は機械に比べてコストがかかるが、検証の結果が機械よりも信頼できる。専門家であるアノテーターによる検証は、この 3 種類の検証方法の中で最もコストがかかるが、検証の結果が最も信頼できる。

学習データとして固有表現と、それに対する、機械・クラウドワーカー・アノテーターによる検証結果が与えられている。この学習データを利用して、問い合わせモデルを学習する。問い合わせモデルは、固有表現を入力とし、問い合わせ先（機械・クラウドワーカー・アノテーターのいずれか）を出力する。

4　既存手法：クラス分類を機械または人間に依頼する

既存手法 [4] はある特徴量 $\boldsymbol{x}$ を n クラスに分類する分類問題において、人間に分類を依頼するか機械学習で分類を行うかを決める手法を提案している。

[4] におけるモデルの概略を図 2 に示す。特徴量 $\boldsymbol{x}$ に対するモデルの出力を $g(\boldsymbol{x})$ とする。モデルの出力は、$n+1$ 次元になっており、$g_i(\boldsymbol{x})(0 \le i \le n-1)$ はモデルの出力 $g(\boldsymbol{x})$ の i 番目の出力で、i 番目のクラスに対するモデルの予測確率に該当する出力になっている。$g_\perp(\boldsymbol{x})$ は n 番目の出力で、人が予測した場合の予測確率に対応している。

$\boldsymbol{x}$ を特徴量を表すベクトル、m は人が行った予測ラベル、y は正解ラベルとする。$\mathcal{Y}$ は予測クラスの集合で、$\perp$ は人に問い合わせを行うか機械学習モデルの予測を採用するかを決定する際に用いる追加のクラスを表す。$\mathbb{I}_z$ は、論理式 z が正しい時 1、誤りであるとき時 0 を返す関数である。

[4] で提案されている損失関数を式 1 に示す。

$$-(\alpha \cdot \mathbb{I}_{m=y} + \mathbb{I}_{m\neq y})\log\left(\frac{\exp(g_y(\boldsymbol{x}))}{\sum_{y'\in\mathcal{Y}\cup\perp}\exp(g_{y'}(\boldsymbol{x}))}\right) - \mathbb{I}_{m=y}\log\left(\frac{\exp(g_\perp(\boldsymbol{x}))}{\sum_{y'\in\mathcal{Y}\cup\perp}\exp(g_{y'}(\boldsymbol{x}))}\right) \tag{1}$$

この式は、クラス分類を行う際に、機械学習モデルの予測を表す関数 $h(\boldsymbol{x})$ と、人にタスクを解く依頼を行うか機械学習モデルの予測を採用するかを決める関数 $r(\boldsymbol{x})$ の学習を同時に行うことを目的としている。$r(\boldsymbol{x})$ の定義を式 2 に示す。$r(\boldsymbol{x})$ は、モデルの予測確率の最大値と人の予測確率を比較し、確率の高い方の予測を採用するようになっている。$r(\boldsymbol{x}) = 1$ の時、人に問い合わせを行い、$r(\boldsymbol{x}) = 0$ の時モデルの予測を採用する。

この式で、人の予測が正解している時に正解ラベル y に対応するモデルの出力 $g_y(\boldsymbol{x})$ と、人に依頼する際に使うラベル $g_\perp(\boldsymbol{x})$ の両方の値が大きくなるように学習を行う。この時、モデルの予測を採用させるように学習させるか、人の予測を採用させるように学習させるかを制御するために、$\alpha(0 \leq \alpha \leq 1)$ を導入している。人の予測が間違っている際、正解ラベルに該当するモデルの出力確率が大きくなるように学習を行う。

$$r(\boldsymbol{x}) = \mathbb{I}_{\max_{y\in\mathcal{Y}} g_y(\boldsymbol{x}) \leq g_\perp(\boldsymbol{x})} \tag{2}$$

5 提案手法：検証を機械・非専門家・専門家にいずれかに依頼する

既存研究では、クラス分類を人間が行うか機械が行うかを決めることを目指していた。本研究では、固有表現候補の検証を、機械、非専門家、専門家のいずれに依頼するかを決めることを目指す。

提案手法における損失関数を式 3 に示す。

$$-(\alpha\mathbb{I}_{a=s=c} + \mathbb{I}_{a=s}) \cdot \log\left(\frac{\exp(g_0(x))}{\sum_{i=0}^{2}\exp(g_i(x))}\right) - ((1-\alpha)\mathbb{I}_{a=s=c} + \mathbb{I}_{a=c}) \cdot \log\left(\frac{\exp(g_0(x))}{\sum_{i=0}^{2}\exp(g_i(x))}\right) - \mathbb{I}_{s\neq a\wedge s\neq c} \cdot \log\left(\frac{\exp(g_2(x))}{\sum_{i=0}^{2}\exp(g_i(x))}\right) \tag{3}$$

機械による検証結果を s、クラウドワーカーによる検証を c、アノテーターによる検証結果を a とする。問い合わせモデルの出力を $g(\boldsymbol{x})$ とする。これは 3 次元のベクトルを表す。問い合わせ先の決定は、$\mathrm{argmax}(g(\boldsymbol{x}))$ として決定する。これは、[4] における関数 $r(\boldsymbol{x})$ に該当し、計算結果が、0 の時機械による検証、1 の時クラウドワーカーに検証を依頼し、2 の時アノテーターに依頼する。$g_i(\boldsymbol{x})$ は問い合わせモデルの出力 $g(\boldsymbol{x})$ の i 番目の値を表している。$g_0(\boldsymbol{x})$ は機械へ問い合わせることが正しい確率、$g_1(\boldsymbol{x})$ はクラウドワーカーへ問い合わせることがが正しい確率、$g_2(\boldsymbol{x})$ はアノテーターへ問い合わせることが正しい確率に対応している。

式 3 の第 1 項では、アノテーターと機械による検証結果が一致している時と、機械とクラウドワーカーの検証結果が正解している時に $g_0(\boldsymbol{x})$ が大きくなるように学習を行う。同様に、第 2 項では、アノテーターとクラウドワーカーの検証結果が一致している時と、機械とクラウドワーカーの問い合わせ先の検証結果が正解している時に $g_1(\boldsymbol{x})$ が大きくなるように学習を行う。また、機械とクラウドワーカーの検証結果が間違っている時にアノテーターに対して問い合わせを行いたいので、$g_2(\boldsymbol{x})$ が大きくなるように学習を行う。α は機械とクラウドワーカーの検証結果が合っている時に、どちらに優先して問い合わせるかを決めるハイパーパラメータである。α の値が 0 に近いとクラウドワーカーによる検証を優先し、1 に近いと機械による検証を優先する。

6 実験

提案手法の式 3 を使って学習させたモデルと、ランダムに問い合わせを決めた際のスコアを比較した。ランダムに問い合わせ先を決める際、モデルの問い合わせ件数を揃え、シード値を 100 回変えてスコアを計測し、その平均値をランダムに問い合わせを決めた際のスコアとした。

6.1 使用する深層学習モデル

最初に、固有表現として抽出されたテキストを RoBERTa[6] を用いて特徴表現に変換する。RoBERTa の事前学習として、森羅 2022 の固有表現抽出タスクのデータセットを用い、固有表現抽出タスクの学習を行った。変換した特徴表現は次元が大きいため、畳み込み層を 2 層使い次元削減を行った。この特徴表現を MLP に入力して 3 次元のベクトルに変換し、問い合わせ先を決定する。ここで用いるモデルのアーキテクチャを図 3 に示す。モデルの学習を行う時は、RoBERTa のパラメータは固定し、畳み込み層と MLP のみ学習を行った。

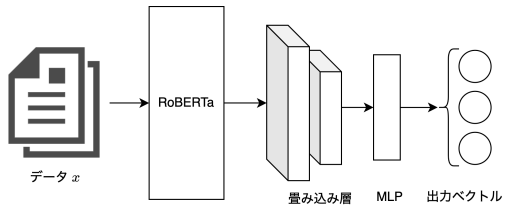

図 3　使用するモデルの概略図

最適化手法には Adam[7]、学習率の大きさは 1×10^{-5}、MLP は 3 層用い、活性化関数には ReLU を用いた。エポック数は 150、バッチサイズは 8 として学習を行った

6.2　データセット

本実験では、人工的なデータを作成しシミュレーションを行った。

データセットの作成に用いたのは、Wikipedia の Lake カテゴリに属するページの中から固有表現として抽出されたテキストを用いた。固有表現抽出には、異なる 6 種類の機械学習モデルを用いている。そのため、同じテキストに対して同じ固有表現が抽出される場合もある。

固有表現を抽出した機械学習モデルの数に閾値を定め、検証結果を決定した。例えば、閾値を 2 に設定した場合 2 つ以上の機械学習モデルが抽出した場合正解、そうではない場合間違いと判定したとする。この操作を閾値を変えて、機械とクラウドワーカーによる判定結果として実験した。機械による判定結果とクラウドワーカーによる判定結果の閾値には、それぞれ 0 と 2 を設定した。

判定結果として、機械の判定結果よりもクラウドワーカーによる判定結果の方が正解率が良くなるようなケースで実験する。そのため、検証結果がアノテーターの解答と機械の検証結果が一致しているもののうち、機械の検証結果を 4 割をランダムに選び、間違ったものにした。

このデータセットの、データ数は 8651 件である。

6.3　結果

α を変化させた時の、問い合わせモデルによる問い合わせ先の回数を集計したところ、α が 0 の時 0 件、α が 1 の時 56 件であった。そのため、α が 0 に近くなると機械に対する問い合わせ回数が増加したことがわかる。また、クラウドワーカーへの問い合わせは α が 0 の時 773 件、α が 1 の時 711 件であったそのため、α が 0 に近くなるとクラウドワーカーへの問い合わせ回数が増加していることがわかった。

	モデルによる振り分け				ランダムに振り分け			
α	Acc.	F1	Rec.	Pre.	Acc.	F1	Rec.	Pre.
0	0.825	0.809	0.799	0.842	0.868	0.753	0.836	0.736
0.1	0.826	0.81	0.8	0.842	0.873	0.757	0.836	0.744
0.2	0.81	0.8	0.843	0.87	0.755	0.836	0.741	0.826
0.3	0.809	0.799	0.842	0.872	0.757	0.836	0.744	0.825
0.4	0.822	0.807	0.797	0.839	0.872	0.757	0.836	0.744
0.5	0.819	0.803	0.794	0.837	0.872	0.757	0.835	0.747
0.6	0.818	0.802	0.793	0.836	0.87	0.753	0.827	0.742
0.7	0.812	0.796	0.788	0.831	0.872	0.75	0.83	0.738
0.8	0.805	0.789	0.781	0.824	0.865	0.75	0.827	0.741
0.9	0.798	0.782	0.775	0.818	0.858	0.736	0.81	0.727
1.0	0.774	0.768	0.812	0.861	0.733	0.802	0.733	0.789

表 1　α の値を変化させた時のスコア

α を変化させた時の、正解率、F1 値、精度、再現率とランダムに問い合わせ先を決定した時のスコアを表 1 に示す。

表 1 より、正解率はランダムに問い合わせを行うよりも良くなっていることがわかる。一方で、F1 値においてはランダムに問い合わせを行った場合よりも悪くなっている。

これより、提案手法で学習させたモデルの方がランダムに問い合わせ先を決めるよりも、検証の精度が良くなっていることがわかる。

7　まとめと今後の展望

本論文では、固有表現抽出モデルが Wikipedia のページから抽出した固有表現とそのラベルが合っているかを検証する方法を選択する手法を提案した。提案した手法は、ランダムに検証する方法を決めるよりも良い正解率を達成したが、F1 値については改善しなかった。

今後は、専門家であるアノテーターに対する問い合わせ回数を制御するためのパラメータを導入し、より精度の良くなる検証ができることを目指す。

参考文献

[1] Satoshi Sekine, Kouta Nakayama, Maya Ando, Yu Usami, Masako Nomoto, and Koji Matsuda. SHINRA2020-ML: Categorizing 30-language Wikipedia into fine-grained NE based on “Resource by Collaborative Contribution” scheme. **In Proceedings of the 3rd conference on the Automated Knowledge Base Construction (AKBC 2021)**, 2021.

[2] Maithra Raghu, Katy Blumer, Greg Corrado, Jon M. Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human

effort. **arXiv preprint**, Vol. abs/1903.12220, , 2019.

[3] Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. Differentiable learning under triage. In **Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual**, pp. 9140–9151, 2021.

[4] Hussein Mozannar and David A. Sontag. Consistent estimators for learning to defer to an expert. In **Proceedings of the 37th International Conference on Machine Learning, ICML 2020**, pp. 7076–7087, 2020.

[5] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In **Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020**, pp. 1526–1533, 2020.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019.

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **3rd International Conference on Learning Representations, ICLR 2015**, 2015.