Pre-Distillation Ensemble: リソース構築タスクのためのアンサンブル手法

中山 功太 †,‡　栗田 修平 ‡　小林 暁雄 ‡　関根 聡 ‡
† 豊橋技術科学大学　‡ 理研 AIP
{kouta.nakayama, shuhei.kurita, akio.kobayashi, satoshi.sekine}@riken.jp

1 はじめに

今日まで、多くの共有タスクが自然言語処理の進歩に大きく貢献してきた。しかし、それらタスクのほとんどは技術開発が目的であり、構築されたリソースの公開は行われない場合が多い。この様な参加者の努力が浪費されている状況は無視できるものではない。近年、タスク参加者による努力をリソース構築といった形で共有する “Resource by Collaborative Contribution(RbCC)” という考え方が考案された。RbCC では更にその後のアンサンブル研究によるリソースの更なる精度向上が期待される。

RbCC に準拠する共有タスクにより得られた複数のリソースに対しアンサンブルを行う場合、以下の様な状況が考えられる。**i)** ヒューリスティックを含む様々なシステムを考慮するため出力確率等は使用できない。**ii)** 多くのラベル無しコーパスに対するシステム予測が利用できる。

実際に RbCC に準拠する共有タスクには “森羅プロジェクト” が挙げられる。これは、Wikipedia 記事からの属性値抽出により知識ベース構築を行うタスクである。その場合以下の様な条件が加わる。**iii)** アノテーションコストが高くアンサンブルに使用可能なデータがない。**iv)** 膨大な選択肢から少数の選択を行うタスクでありシステム間の合意が発散しやすくアンサンブルが困難である。以上は限定的な状況ではあるが、これは今後行われる様々なリソース構築タスクにも当てはまり得る状況である。しかしながら、現在のアンサンブル手法はその様な状況で最善に機能するとは言い難い。そのため、以下の様な特徴を持つ手法の考案が望まれる。i) 出力確率を疑似的に推定可能である。ii) ラベル無しコーパスに対する予測を活用できる。iii) 教師なしアンサンブルに近い状況で動作する。iv) システム間の合意に強く依存しない。本研究では、以上の特徴を満たす、様々なリソース構築共有タスクに一般化されたアンサンブル手法の考案を行い、森羅プロジェクトで得られた複数のリソースに対するアンサンブル結果によりその優位性を示す。また本研究により得られた結果は知識ベースとして公開[1]する。

2 森羅プロジェクト 2019

森羅プロジェクトでは、拡張固有表現階層 [1] で定義された約 200 カテゴリーに分類済みの Wikipedia 記事から、同階層により定義された属性に対応する値の抽出を行う。値には文章中の位置情報を付与する必要がある。そのため、表層の文字列が同等であっても出現する全ての位置を特定する必要がある。参加者は、学習用データと参加するカテゴリーに属する記事全てが配布され、その記事全てに対するシステムの予測結果のみを提出する必要がある。

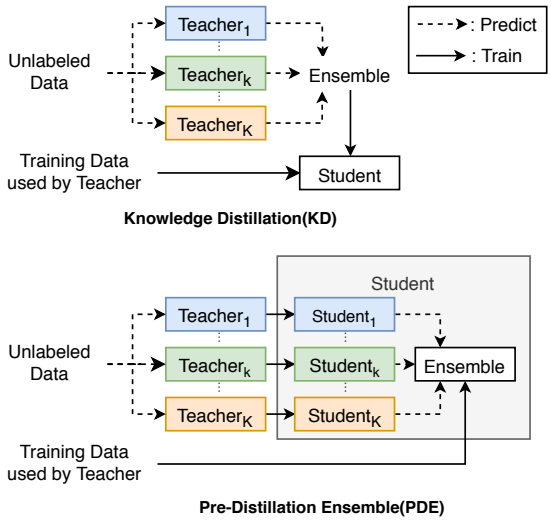

図 1: Knowledge Distillation(KD) と Pre-Distillation Ensemble(PDE) フレームワーク

森羅プロジェクト 2019 では 5 カテゴリーを対象とする JP-5 タスクと 30 カテゴリーを対象とする JP-30 タスクが行われた。この内 JP-30 タスクは、“地名” と “組織名” のクラスに分けられる。森羅プロジェクト 2019 は 2019 年 4 月から 9 月にかけて実施され、9 システムの結果が提出、公開されている。本研究ではこれらデータに対してアンサンブルを行うが、JP-30 タスクの結果の内 28 カテゴリーのみ対してアンサンブルを行う[2]。アンサンブルには、各システムの結果に加えて、タスク参加者に配布されたカテゴリーごと 158〜200 件記事分の学習データと 227〜12008 件の記事が利用できる。また、開発用データとして”地名”クラスの”湖沼名”カテゴリーにおいて 100 件分の正答データが利用できる。

[1] http://shinra-project.info/download/

[2] JP-5 タスクでは人為的なサンプリングにより、記事全体と学習データ、評価データの範囲の分布が異なるため本研究では対象外としている。また、評価データの有無の関係により、JP-30 から 2 カテゴリーは対象外としている。

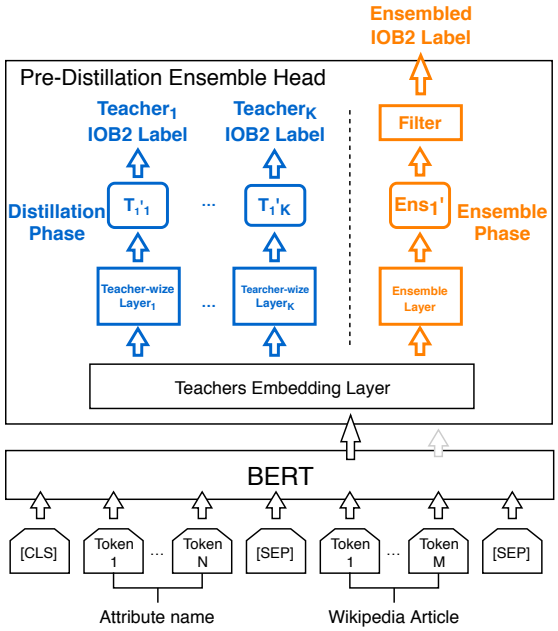

図 2: PDE モデル

3 関連研究

アンサンブル専用の正答データを利用できない点で、本研究は教師なしアンサンブルに近いものである。教師なしアンサンブルには、多数決や加重平均、DS 法 [2] などがある。加重平均には、各システムの出力確率等を用いるが、本研究の様に各値の出力確率が利用できない場合は、多数決と同意である。DS 法は各値に対するラベルの確率と各モデルのエラー率を EM アルゴリズムにより推定する手法である。しかし各システム間の合意が発散しやすいタスクでは、以上の様なシステムの合意に依存する手法は効果的に機能しない可能性がある。

本研究と同様に共有タスクの結果に対してアンサンブルを行った研究は、KBP の Slot Filling Validation/Ensemble タスク [3, 4] において多く見られる。しかしこれらの研究は、昨年のタスクに共通して参加しているシステムのスコアや、各値の正誤分類を人手により行ったデータを利用可能である点で、教師ありアンサンブルである。そのため、アンサンブルのための正答データが利用できない本研究とは異なる。

また次章で解説する提案手法は図 1 上部に示す様な Knowledge Distillation(KD)[5, 6] から着想を得ている。KD は複数モデルのアンサンブル結果、もしくは膨大なパラメーターを持つモデルの結果から新規モデルを学習することで、モデルの推論速度の向上や必要パラメーター数の削減を行うことを目的とした手法である。この際、学習元となるモデルを教師 (Teacher) モデル、結果から新規学習を行うモデルを生徒 (Student) モデルと呼ぶ。また、モデル出力を使用して新規モデルを学習することを蒸留 (Distillation) と呼ぶ。KD は自然言語処理においても多くのタスク [7, 8] で使用されている。

クラス	最良システム	多数決	DS 法	BERT	提案手法
地名	61.94	62.39	57.84	59.54	**68.76**
向上	-	+0.45	-4.11	-2.41	**+6.82**
組織名	62.93	56.99	51.41	55.01	**66.41**
向上	-	-5.94	-11.52	-7.93	**+3.48**
全体	62.56	59.08	53.93	56.73	**67.30**
向上	-	-3.49	-8.63	-5.83	**+4.74**

表 1: クラスごとのアンサンブル結果

4 Pre-Distillation Ensemble

本研究では 1 章で列挙した様な特徴を満たすアンサンブル手法である **Pre-Distillation Ensemble(PDE)** を提案する。図 1 下部に PDE のフレームワークを示す。これは、図 1 上部の Knowledge Distillation(KD)[5, 6] から着想を得たものである。PDE では、教師モデルの結果を単一の生徒モデルを用いて蒸留した後、教師モデルで使用された学習データを再利用してモデル内部においてアンサンブルを行う。これにより以下の様な利点がある。**i) 蒸留による教師システムの出力確率等の情報復元が期待され、特徴量としての活用が期待できる。ii) ラベル無しコーパスに対する教師モデルの予測が活用できる。iii) アンサンブル専用の学習データを必要としない。iv) 各教師モデル間の合意に強く依存しない。**

図 2 に、本研究で使用する PDE モデルを示す。これは BERT[9] と **Pre-Distillation Ensemble Head(PDE Head)** で構成される。学習は蒸留段階 (Distillation Phase) とアンサンブル段階 (Ensemble Phase) に分けられる。

森羅プロジェクトタスクを各トークンに対する IOB2 タグ予測と考える。BERT への入力として与えられた属性名と Wikipedia 記事は、各トークンごとにベクトルに変換され、共通の PDE Head に与えられる。蒸留段階では、トークンベクトルは共通の Teachers Embedding Layer(TEmL) を経て、教師システムごとに定義された Teacher-wise Layer(TwL) へと渡され、各システムの出力した IOB2 タグを予測する。アンサンブル段階では、TEmL を経て、Ensemble Layer(EnL) へと渡され、各教師システムの学習に使用された学習データの IOB2 タグを予測する。TEmL、TwL、EnL はそれぞれ 1 層の全結合層である。また、評価時にのみ PDE-Head の出力にフィルターを適用する。フィルターは教師システムの出力との積集合である。

5 実験

5.1 BERT 事前学習

PDE モデルに使用するため BERT の事前学習を行う。モデルサイズには *base* サイズ [9] を使用し、学習

クラス	カテゴリー	学習データ記事数	蒸留用記事数	最良システム	多数決	DS 法	BERT	提案手法
地名	GPE その他	200	395	56.45	51.53	48.82	58.26	**64.15**
	国内地域名	200	2000	50.49	50.72	39.84	45.85	**53.73**
	国名	158	1304	64.26	63.67	59.96	60.34	**70.99**
	地名 その他	200	2000	49.00	52.91	38.61	45.56	**56.53**
	地形名 その他	200	2000	62.65	60.53	56.27	51.41	**66.38**
	大陸地域名	147	269	56.38	56.36	54.29	55.17	**63.15**
	山地名	200	2000	62.73	65.15	63.50	62.47	**67.65**
	島名	173	2000	67.40	69.13	63.47	62.57	**75.29**
	河川名	200	2000	64.92	63.88	59.17	64.94	**72.50**
	海洋名	200	291	65.65	65.91	63.27	61.42	**72.10**
	温泉名	190	1080	74.24	**79.02**	74.82	72.60	78.93
	湖沼名	200	772	63.09	61.26	58.33	64.92	**71.02**
	湾名	200	354	67.47	65.39	62.43	58.74	**75.20**
	都道府県州郡名	198	2000	67.25	70.83	67.18	67.14	**76.05**
組織名	企業グループ名	200	386	65.03	59.97	50.02	52.56	**69.60**
	公演組織名	196	2000	71.43	68.85	65.52	64.66	**75.19**
	国際組織名	191	949	52.71	53.23	44.30	44.54	**60.15**
	家系名	200	1905	69.66	65.96	68.11	66.98	**75.23**
	政党名	199	1543	52.24	52.68	44.92	47.28	**57.87**
	政府組織名	200	2000	51.20	57.33	51.52	53.36	**62.33**
	政治的組織名 その他	200	1177	**47.55**	36.88	31.20	42.27	47.02
	民族名	200	1133	56.71	52.57	46.04	55.37	**61.76**
	競技リーグ名	189	841	**63.86**	44.30	34.75	49.00	61.25
	競技団体名	199	2000	54.92	55.03	49.61	54.74	**62.35**
	競技連盟名	200	790	56.94	57.95	54.52	41.67	**60.43**
	組織名 その他	183	2000	53.95	54.58	47.71	52.21	**60.65**
	軍隊名	200	2000	67.52	54.85	48.44	57.48	**69.11**
	非営利団体名	200	2000	59.75	53.48	45.19	46.41	**60.84**

表 2: カテゴリーごとのアンサンブル結果

は RoBERTa[10] と同様の目的関数で行う。学習コーパスには日本語 Wikipedia を使用する。

BERT への入力文章は事前に、Janome[3]を使用して単語分割され、BPE[11] を使用してサブワード分割[4]される。

5.2 蒸留段階 (Distillation Phase)

PDE の学習はカテゴリーごとに行う。蒸留に使用する教師システムの予測は、最大 2000 件に制限され、それら予測は学習データの範囲に対する予測を全て含む様選択される。選択された蒸留用データは、90%が学習用データ、10%が開発用データとして使用される。学習は開発データによるスコアが 4 エポック以上更新されない場合、もしくは 30 エポックを超えた場合に終了し、スコアが最良なモデルを選択する。

5.3 アンサンブル段階 (Ensemble Phase)

アンサンブル段階では教師モデルの学習に使用されたデータを再利用する。データ分割や学習の終了条件は蒸留段階と同様である。学習パラメータは、アンサンブル段階の開発データのスコアを最大化する様選択される。このパラメータを決定する操作は “湖沼名” カテゴリーのみで行う。

5.4 アンサンブル結果

提案手法の優位性を示すため以下の様な手法と比較する。

最良システム 教師システムのうち最良のスコア。
多数決 値に対する合意数を使用した多数決。表が拮抗した場合はランダムで選択。
DS 法 [2] 3 章参照。
BERT 森羅プロジェクト参加者と同条件において BERT を学習。

クラスごとのアンサンブル結果を表 1 に示す。値は全て F1 値のマイクロ平均であり、向上は最良システムからの差分を示している。提案手法は、他の手法と比較し大きな向上を得ていることが確認でき、これは手法の優位性を示すものである。多数決や DS 法では、最良システムと比較するとほとんどの場合で大きなスコアの低下が確認される。これは両者がシステム間の合意に大きく依存する手法であり、本タスクの様な各システム間の合意が分散しやすいタスクでは有効に機能しないためであると考えらえる。また、提案手法と BERT を比較すると非常に大きい F 値の向上が確認できる。これは、蒸留が効果的に機能していることを示している。

カテゴリーごとのアンサンブル結果を表 2 に示す。また、教師データの学習と蒸留に使用した記事数も同時に示す。各スコアは、全て F1 値のマイクロ平均である。提案手法は、28 カテゴリー中 25 カテゴリーにおいて最良の結果であることが確認できる。これは提案手法の安定性を示している。“温泉名” カテゴリーでは、提案手法は多数決に対し僅差で劣っている。最良

[3]https://mocobeta.github.io/janome/
[4]日本語 Wikipedia を使用して学習、マージ数は 10000 に設定。

手法	精度	再現率	F1
提案手法	74.29	68.51	**71.28**
フィルターの切除	70.30	68.93	69.61
ラベル無しコーパスの切除	79.34	59.83	68.22
多数決	86.65	46.92	60.87
Knowledge Distillation	85.06	48.49	61.77
最良システム	61.14	65.26	63.13

表 3: アブレーション試験結果

システムのスコアから、“温泉名” は比較的簡単なカテゴリーであったことが推測できる。そのため、システム間の同意が集中しやすく多数決の方が優位なスコアであったと考えられる。しかし、各システムの結果を事前に把握することは不可能である上、提案手法は最良システムと比較し非常に優位なスコアである。“政治的組織名 その他” と “競技リーグ名” では、最良システムと比較して提案手法によるスコアの低下が見られる。しかし、システムのスコアの算出と最良システムを選ぶ際に使用する評価データは同じであるため、最良システムは評価データに対して過適合している。そのため、直接には比較できないことに注意する必要がある。また、2 カテゴリーでは提案手法は他手法よりは良いスコアである。

5.5 アブレーション試験

提案手法の各機構の効果を示すため、アブレーション試験を行った結果を表 3 に示す。本試験は “湖沼名” カテゴリーに用意された開発データを使用する。以下でそれぞれの説明を行う。

フィルターの切除 本研究で使用した PDE モデルは評価時にフィルターにより、教師モデルが出力していない値を排除している。これは、生徒モデルによる教師モデルのエラーの増幅を防ぐためである。フィルターを切除した結果から、精度の大きな低下が確認でき、これはフィルターが効果的に機能していることを示している。

ラベル無しコーパスの除去 ラベル無しコーパスに対する教師モデルの予測を除去した結果、精度は大きく向上するが再現率が大きく低下し、最終的に F1 値において低下が確認される。これは、ラベル無しコーパスに対する予測が利用できるタスクの利点を示している。しかし、切除後であっても最良システムの F1 値と比較して非常に良いスコアであることから、ラベル無しコーパスに対する予測は PDE に必須でない可能性を示唆している。

Knowledge Distillation(KD) KD では、PDE と異なり蒸留の前にアンサンブルを行っている。そのため、スコアはアンサンブル結果に強く依存すると考えられる。はじめに多数決の結果に注目すると、精度のみが非常に高いことが確認できる。そのため、KD においても同様な結果が確認される。しかし、多数決と比較すると KD では F1 値が向上しており、精度へのバイアスの蒸留による緩和が見られる。だが、PDE と比較するとかなり劣る結果であることから、合意ベースの手法が効果的に機能しないタスクの場合、PDE は非常に有効であることが分かる。

6 おわりに

本研究では、森羅プロジェクト 2019 で得られた複数の知識ベースに対し最適にアンサンブルを行うことのできる Pre-Distillation Ensemble(PDE) を提案し、実験におけるスコアの向上によりその優位性を示した。この提案手法は共同貢献によるリソース構築を目的とした “RbCC” に準拠した多くのタスクで活用できると考えられる。そのため、今後タスク参加者の努力を共有するためにも、複数の共有タスクが “RbCC” に準拠していくことを期待する。

参考文献

[1] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In *LREC*, 2002.

[2] A. Philip Dawid and Allan Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. 1979.

[3] Mihai Surdeanu. Overview of the tac2013 knowledge base population evaluation: English slot filling and temporal slot filling. In *Proceedings of the Sixth Text Analysis Conference (TAC 2013)*, 2013.

[4] Mihai Surdeanu and Heng Ji. Overview of the english slot filling track at the tac2014 knowledge base population evaluation. In *Proceedings of the Seventh Text Analysis Conference (TAC2014)*, 2014.

[5] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pp. 2654–2662. Curran Associates, Inc., 2014.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[7] Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, and Noah A. Smith. Distilling an ensemble of greedy dependency parsers into one MST parser. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1744–1753, Austin, Texas, November 2016. Association for Computational Linguistics.

[8] Markus Freitag, Yaser Al-Onaizan, and Baskaran Sankaran. Ensemble distillation for neural machine translation. *CoRR*, Vol. abs/1702.01802, , 2017.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019.

[10] 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. *CoRR*, 2019.

[11] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.