共有タスクにおけるGA重み付け加重投票を用いた属性値アンサンブル

中山 功太[†,‡], 小林 暁雄[†], and 関根 聡[†]
[†] 理研 AIP
[‡] 豊橋技科大 博士前期課程 情報・知能工学専攻
{kouta.nakayama ,akio.kobayashi ,satoshi.sekine}@riken.jp

1 はじめに

機械可読に構築された知識ベースは多くのシステムに必要不可欠である。説明可能なNLPシステムにおいてもその限りではない。インターネットの普及とともにオンライン上には数々の知識ベースが作成された。その中でもWikipediaは多くの固有名詞を扱っており、非常に貴重な知識資源である。しかし、Wikipediaは人間が読むため作成されており、機械可読な構築ではない。Wikipediaを機械可読に再構築する試みは多数存在し、その中で構成された知識ベースにDBpedia、Freebase、YAGO、Wikidataなどがある。だが、これらはトップダウン式でない設計や、クラウドワーカーによる作成により、カバレージの低さや信頼性の低さといった問題に悩まされている。また、テキストから知識の自動構築を行うシステムを開発するKBPやCoNLLといったプロジェクトも存在する。しかし、これらは最終的な知識構築が目的でなく、参加システムによる出力結果はリソースとして共有されない。これら問題に対処すべく、Wikipediaからの知識構築タスクとして開始されたプロジェクトが、森羅プロジェクト[1]である。本プロジェクトは、“Resource by Collaborative Contribution(RbCC)”の考えに基づいており、参加システムの結果は共有され、知識ベースとして利用可能であるとともに、その後のアンサンブルによるさらに洗練した知識ベースの構築を行うことが可能である。本論文の目的は、森羅プロジェクトにより共有されたデータに対して、GA重み付け加重投票を用いたアンサンブルを行い、その優位性を評価することであり、またその結果によって“RbCC”の考え方の元行われる森羅プロジェクトの優位性を示すことである。

2 森羅プロジェクト

森羅プロジェクトは“RbCC”の考え方の元、Wikipedia記事から拡張固有表現階層[7]に従い、知識構築を行うタスクである。Wikipedia各記事は、最新の拡張固有表現階層[7, 6]により定義された200のカテゴリに分類されており、本タスクの目的は、対応したカテゴリに割り当てられた属性の値の抽出である。参加者は、トレーニングデータに基づく抽出システムの開発と、対象カテゴリに含まれるWikipedia記事全件の抽出結果の提出が要求され、例えば、“人名”カテゴリでは、20万件の記事に対する抽出が必要である。森羅2018[10]は、2018年4月より10月まで実施され、対象カテゴリは“人名”、“企業”、“市区町村”、“化合物”、“空港”であった。また配布されたトレーニングデータは600件であり、8グループが参加、16のシステムが結果を提出した。

3 関連研究

アンサンブルは、システムや結果の統合を行う手法の総称であり、自然言語処理においても様々な分野において使用される。アンサンブルがよく使用される分野には複合語判別(CWI)があり、主に共有タスクにおいて多くの使用が報告されている。その中でも、良い成績を残しているものが、ソフト投票[4]やハード投票[8]である。しかし、これら手法は値の信頼度が必要であり、値に対して信頼度を持たない森羅2018のデータでの使用は不可能である。その中で、森羅2018において使用できる手法に多数決による投票[1]があり、比較のため実装を行う。また、機械学習において使用されるアンサンブルには、ブースティング[5]、バギング[2]、スタッキング[9]等がある。ブースティングやバギングは複数の学習器を用いて、各学習器よりも優れた結果を残す手法であり、対して、スタッキングは各システムの出力に対してメタ学習器によりアンサンブルを行う手法である。参加システムの出力結果のみが使用できる森羅2018ではスタッキングのみ利用可能である。KBP内のスロットフィリング(SF)において、スタッキングを採用したシステムに、[3]があり、SFシステムの信頼度を特徴量として使用し、SVMに

[1] https://aip.riken.jp/labs/goalorient_tech/lang_inf_access_tech/森羅：wikipedia構造化プロジェクト2018/?lang=ja

よる統合を行なっている。しかし SVM を使用するため、最大化の対象は最終的に構築されるデータベースの F 値ではない。対象的に、本研究では GA アルゴリズムを用いて重みを最適化した、加重投票を用いており、最大化の対象は最終的に構築されるデータベースの F 値となる。また、最適化手法により重み付けされた加重投票を用いることにより、同様のアプローチを用いたシステムが多数参加することによる、最終的な出力への影響を軽減することが可能である。

4 手法

本実験では、“GA 重み付け加重投票” の実装に加え、比較のため、“単純投票”、“加重投票” の実装も行なう。以下で手法の説明を行う。また、全ての手法において、開発用データとテスト用データが必要であるため、森羅参加者に対して配布されていない計 100 件のデータが 2 等分して用いられ、結果の信頼度を高めるため交差検証が適用される。

4.1 単純投票

本手法では、抽出に参加するシステムのうち、ある属性に対して値 v を出力するシステムが n システム存在した場合、$n > t$ であれば、その値の採用を決定する。ここで使用される閾値 t は開発データを用いて決定され、最大化の対象は F 値である。

4.2 加重投票

優秀なシステムには、他のシステムに比べより大きな重みが用いられるべきと考えるのは当然である。そのため、この手法において用いられる重みは、開発用データにより計測された各システムの精度である。単純投票における、システム数 n に対して、値 v を出力するシステムの精度の合計が用いられ、また、閾値 t の決定は単純投票と同様である。

4.3 GA 重み付け加重投票

加重投票では、最適な重みの割り当てを行うことで、より洗練されたアンサンブルが可能である。最適な重みは、出力されるアンサンブル結果の F 値を最大化することにより、決定されるべきであり、本研究ではこの最適な重みの探索に、遺伝的アルゴリズム (GA) を用いる。GA は生物の遺伝をシミュレートした最適解探索法であり、ランダムに生成された遺伝子集合に対し、選択、交差、突然変異などによって適応度の高い遺伝子を生存させる手法である。今回最適化する重みは連続値であり、そのため使用する交叉法には、BLX-α を用いる (BLX-α を使用する場合には、突然変異を行う必要はない)。また、世代交代法に用いるのは、一般的に使用されている MGG である。

カテゴリ	属性数	インスタンス数
空港	24	2061
市区町村	26	2699
企業	33	3066
化合物	15	2181
人名	22	2839
合計	120	12846

表 1: カテゴリ毎の属性数とデータ 100 件中に含まれるインスタンス数

5 実験

5.1 使用データ

実験と評価には使用するデータは、森羅プロジェクト 2018 において参加者に配布されていない計 100 件の正答データであり、表 1 はその概要である。

5.2 ベースラインシステム

比較のためのベースラインシステムは、各カテゴリにおいて F 値が最良のシステムによる構成される。全てのカテゴリに一貫して単一システムを用いた場合より、優れた結果であり、これはすでに一種のアンサンブルシステムである。しかし、本研究では、単一のシステムに対する改善を示すために、カテゴリごとに最良のシステムの選択を行う。

6 結果

6.1 アンサンブル結果

表 2 は、ベースラインシステムと 3 つのアンサンブル手法の F 値であり、向上値はベースラインシステムに対するアンサンブルメソッドの相対的な向上である。単純投票および加重投票による向上値は、それぞれ 4.7 および 9.8 である。対して、GA システムによる向上値は、11.4 であり、これは比較手法に対し、かなり有意であると言える。また、アンサンブルによる、これほどまでに大きな F 値の向上は、従来の共有タスクの方式に対して “RbCC” 方式を採用することの大きな優位性を示している。

表 3 は、カテゴリごとの結果である。“化合物” カテゴリにおける向上は、17.7 と最大であるが、最小でも “人名” カテゴリにおいて 4.0 の向上が見られる。深層学習を用いたシステムの改善非常に重要であるが、実際にはアンサンブルを行うことによって、それと同

手法	精度	再現率	F 値	向上値
Baseline	57.4	46.6	51.4	-
単純投票	55.7	56.6	56.1	+4.7
加重投票	**69.5**	54.9	61.2	+9.8
GA 重み付け加重投票	67.6	**58.3**	**62.7**	+11.3

表 2: ベースラインとアンサンブル 3 手法の結果

カテゴリ	ベースライン	単純投票	加重投票	GA 重み付け加重投票	向上値
空港	71.8	80.9	84.9	**87.0**	+15.2
市区町村	46.0	54.6	57.2	**58.4**	+12.4
企業	53.4	51.2	58.3	**60.8**	+7.4
化合物	47.1	56.5	62.9	**64.8**	+17.7
人名	43.9	43.7	47.5	**47.9**	+4.0
平均	51.4	56.1	61.2	**62.7**	+11.3

表 3: カテゴリ毎の結果

等、もしくはそれ以上の改善を得られることがこの実験において、明らかになった。

6.2 システムによる貢献

なぜここまで大きな向上がアンサンブルにより得られたのかを理解することは大切であり、そのために、結果に対して詳細な分析を行う必要がある。表 4 は参加システムが多い “企業カテゴリ” において各システムの貢献度を分析した結果である。この貢献度は、全てのシステムを含むアンサンブル結果と、対象のシステム群を含まないアンサンブル結果の差により求められる。表 4 よりシステム 7,16 の貢献度はそれぞれ 4.0 と 2.2 であり、この 2 つのシステムにより大きな向上がもたらされていることがわかる。しかし、貢献度を使用する場合は、システムの貢献が重複している可能性を考慮する必要がある。例えば、単独での貢献度が 0 に近いシステムであっても、システム 12,13,15 のように合算では 0.8 と良い貢献を示す場合があるためである。また、システムは 1,3,6,14 は人手により作成されたパターンを用いたシステムであり、これらシステム群の貢献度は 2.9 である。これは、システム 7 の次に大きな貢献となり、パターンを用いたシステムが深層学習を用いたシステムに対し非常に補完的に働いていることを示す重要な分析である。全体のシステムにパターンベースのような様々な手法が含まれるのは、“RbCC” の考え方を元にした共有タスクによってもたらされる利点であり、パターンベースのシステムによってもたらされる結果の向上は、“RbCC” の優位性をより裏付けるものである。

システム	+精度	+再現率	+F 値
1 (パターン)	0.9	0.2	0.4
4 (パターン /ヒューリスティック)	1.3	0.2	0.5
5 (深層学習)	0.1	0.0	0.1
6 (パターン)	0.1	0.2	0.1
7 (DrQA)	1.3	5.5	4.0
12 (深層学習)	-0.2	0.1	0.0
13 (深層学習)	0.1	-0.2	-0.1
14 (パターン)	0.1	0.1	0.1
15 (パターン /深層学習)	-0.1	0.0	0.0
16 (深層学習)	3.3	1.7	2.2
1,4,6,14	2.6	3.2	2.9
12,13,15	2.5	-0.2	0.8

表 4: “企業” カテゴリにおける各システムの貢献度

7 考察

実験結果は、アンサンブルシステムによるとても有意な向上と、パターンベースのシステムによる大きな貢献を示したものであった。しかしこの結果は “RbCC” 方式の優位性を示すには十分であるが、アンサンブルシステムの完全性を示すには不十分であった。なぜならアンサンブルシステムの理想値は、精度が 100%であり、また再現率は全てのシステムの結果の和集合の再現率であるが、提案手法は依然としてこの数値に到達していないからである。表 5 はシステムの結果の和集合の再現率と提案手法の再現率との差を計測した損失率である。“空港” カテゴリは損失率は 9.0%と低く、また、精度も 90.6 とこのカテゴリにおいては提案手法がうまく機能している事がわかる。しかし、その他のカテゴリでは、精度は 56.6 から 72.2 であり、損失率は 13.6%から 23.8%と、理想値に対して大きな差がある結果となった。これは、“空港” カテゴリでは、各システムの出力する正答値はかなり一致していたが、その他のカテゴリにおいては、少数 (もしくは単一) のシステムのみでしか出力されない正答値が多く存在し投票において採用されなかったためであると考えられる。このような正答値の採用のため、さらなる改善が必要である。

8 今後について

7 章では、アンサンブルについて改善の余地がある事が示された。この改善には、属性値を出力したシステム情報のみを入力としている現在のアンサンブルシス

カテゴリ	和集合	GA 重み付け加重投票		損失率
	再現率	精度	再現率	
空港	92.7	90.6	83.7	9.0%
市区町村	73.9	56.6	60.4	13.5%
企業	72.5	68.5	54.8	17.7%
化合物	72.4	72.2	58.8	13.6%
人名	65.2	58.0	41.4	23.8%
平均	74.4	67.6	58.3	16.0%

表 5: 各カテゴリ毎の損失率

テムに対し、特徴量を付与することが考えられる。例えば、単語埋め込みや、出現セクション、抽出難易度などである。また、森羅 2018 は値の信頼度や抽出位置の提出は必要でないため共有されたデータ内に含まれていない。しかし、将来的に、共有されるデータにこれら情報が追加された場合、これもまた特徴量として用いる事ができる。また、アンサンブルにより構築された知識ベースの利用手段によっては、現在 F 値を用いている最大化の対象を変更する事が考えられる。例えば、知識ベースに対し、クラウドソーシングを用いて、クリーンアップを行う場合には、現在の F 値の最大化より、より、再現率に対し重みを高く設定した評価関数の最大化を行なった方が良い。これは、クラウドソーシングは膨大なテキスト内からの値の抽出、つまり再現率の向上には向いておらす、値の正誤判定のような、精度の向上させるタスクに対して有効だからである。そして、人手により作成されたパターンのような手法が、アンサンブル後の知識ベースに対して、貢献を残していたことから、さらに異なるタイプのシステムの開発は、更なるアンサンブル後の知識ベースの洗練を行える可能性を提示している。そのため、我々は様々な手法を用いたシステムの開発が必要である。

9 終わりに

本研究の目的は、森羅プロジェクトに基づくアンサンブル手法である GA 重み付け加重投票の提案と、アンサンブルにより構築された知識ベースから “RbCC” 手法の優位性を示すことであった。森羅 2018 によって共有されたデータに対し提案手法によりアンサンブルを行なった結果、平均で 11.3 の F 値の向上が見られ、カテゴリごとでは最大 17.7 の向上が見られた。また、結果に対する分析により、アンサンブル後の結果に対して、パターンベースのシステムが補完的に働いていることが確認された。提案手法には更なる改善点が残されているものの、これらの結果は、“RbCC” の元で行われる知識ベース構築タスクによる優位性を示すには十分であった。“RbCC” の信念の元で、システムの出力データを共有し、アンサンブルによる改善を示す機会を与えてくれた森羅プロジェクトとその参加者に感謝するとともに、知識構築以外のタスクにおいても、“RbCC” の考え方が普及することを願う。

参考文献

[1] David Alfter and Ildikó Pilán. Sb@gu at the complex word identification 2018 shared task. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 315–321. Association for Computational Linguistics, 2018.

[2] Leo Breiman. Bagging predictors. *Machine Learning*, Vol. 24, No. 2, pp. 123–140, Aug 1996.

[3] Nazneen Fatema Rajani, Vidhoon Viswanathan, Yinon Bentor, and R.J. Mooney. Stacked ensembles of information extractors for knowledge-base population. Vol. 1, pp. 177–187, 01 2015.

[4] Gustavo Paetzold and Lucia Specia. Sv000gg at semeval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 969–974. Association for Computational Linguistics, 2016.

[5] Robert E. Schapire. The strength of weak learnability. *Machine Learning*, Vol. 5, No. 2, pp. 197–227, Jun 1990.

[6] Satoshi Sekine. Extended named entity ontology with attribute information. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

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

[8] Nikhil Wani, Sandeep Mathias, Jayashree Aanand Gajjam, and Pushpak Bhattacharyya. The whole is greater than the sum of its parts: Towards the effectiveness of voting ensemble classifiers for complex word identification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 200–205. Association for Computational Linguistics, 2018.

[9] David H. Wolpert. Stacked generalization. *Neural Networks*, Vol. 5, pp. 241–259, 1992.

[10] 関根聡, 小林暁雄, 安藤まや. Wikipedia 構造化プロジェクト「森羅 2018」. 言語処理学会第 25 回年次大会, 2019.