固有表現抽出におけるタグセットの相互適応

鈴木雅也　古宮嘉那子　佐々木稔　新納浩幸

茨城大学工学部情報工学科

{13t4038a, kanako.komiya.nlp, minoru.sasaki.01,
hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

1 はじめに

近年, 古典的な教師あり学習の枠組みを超えた新たな学習手法についての研究が盛んに行われている. そのひとつとして, マルチタスク学習が挙げられる. マルチタスク学習とは転移学習の一種であり, ふたつのタスクについて, 相手のタスクからの知識の転移を相互に行うことで, 双方のタスクの精度向上を目指す学習手法である [1]. 本稿では, 素性に対して異なるタスクのタグセットを追加する方式のマルチタスク学習による学習手法 (タグセットの相互適応) の提案と既存手法との比較による考察を行う. なお, 本研究では, 提案手法を適用するタスクとして, 固有表現抽出タスクを選定した.

2 関連研究

本稿では, 固有表現抽出についてマルチタスク学習を行う. そこで, まず転移学習の関連研究について述べ, 次に固有表現抽出の関連研究について述べる.

神蔦 [2] は, データのどの部分を転移時に変換するか, 転移のどの段階でデータを変換するか, というふたつの基準から, 転移学習を特徴ベース・分離型, 特徴ベース・統合型, 事例ベース・分離型, 事例ベース・統合型の 4 タイプに分類した. 提案手法は, 転移元の素性の中で転移先でも有用なものを選択し, 知識を転移させてから学習を行うため, 特徴ベース・分離型に属する. 特徴ベース・分離型の転移学習に関する先行研究としては, 簡便な帰納転移学習の手法を提案した Daumé[3] や, 同手法を用いた後, フィルタリングを行った Komiya[4] が挙げられる. しかし, これらの研究はマルチタスク学習ではなく, 異なったジャンルの訓練事例を用いて同一タスクの学習を行う領域適応の研究である. また, ニューラルネットワークを用いた古典的なマルチタスク学習の研究に Thrun[5] と Caruana[6] がある. しかし, 我々が知る限り, 素性に対して異なるタスクのタグセットを追加する方式のマルチタスク学習による学習手法に関する論文は存在しない.

本稿ではマルチタスク学習の対象タスクとして, 2 種類の固有表現抽出タスクを用いる. 固有表現抽出とは, 固有名詞に時間や数値といった表現を加えた概念である固有表現を文章中から抽出するタスクであり, 昔から研究が行われてきた. このタスクに関する先行研究としては, 次のようなものが挙げられる. アノテーション手法に関して, 鈴木ら [7] は非専門家による固有表現抽出のタスクとしてのアノテーションを題材に, 既存の固有表現抽出器によるアノテーション結果に対し, 人手で修正を行うアノテーション手法を提案した. また, タグの定義に関して, Sekine ら [8] は Information Retrieval and Extraction Exercise (IREX)[1] で固有表現抽出の共通タスクを行うため, 9 種類のタグを定義した. さらに, Sekine ら [9, 10, 11] は, IREX での定義と Automatic Content Extraction (ACE)[2] での定義を元に, 200 種類の固有表現タグからなる関根の拡張固有表現階層を定義した. そして, コーパスに関して, Iwakura ら [12] は現代日本語書き言葉均衡コーパス (BCCWJ)[13][3] に対し, IREX での定義のタグを付与した BCCWJ NE コーパス[4] を作成した. 橋本ら [14, 15] は CD-毎日新聞'95 データ集[5]や BCCWJ に対し, 関根の拡張固有表現階層のタグを付与した拡張固有表現タグ付きコーパス[6]を作成した.

[1] http://nlp.cs.nyu.edu/irex/index-j.html
[2] https://www.ldc.upenn.edu/collaborations/past-projects/ace
[3] http://pj.ninjal.ac.jp/corpus_center/bccwj/
[4] https://sites.google.com/site/projectnextnlpne/
[5] http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html
[6] http://www.gsk.or.jp/catalog/gsk2014-a/

表 1: 提案手法における訓練事例の素性一覧

転移元	転移先	素性	正解ラベル
α	β	$a_1\ a_2\ ....\ a_n\ l_\alpha$	l_β
β	α	$b_1\ b_2\ ....\ b_m\ l_\beta$	l_α

3 タグセットの相互適応

本稿では, マルチタスク学習のひとつとして, タグセットの相互適応を行う. まず, 類似のふたつのタスク α, β があると仮定する. この際, α, β のタグセットは十分に関連性があるが, 同一のものではないとする. また, α, β がそれぞれ素性 $a_1, a_2, ..., a_n$, 及び, $b_1, b_2, ..., b_m$ を用いており, タスク t の正解ラベルを l_t とすると, 表 1 で示されたような 2 種類の訓練事例が生成される. 本手法では, これらの素性を用いて既存の分類器で学習を行う.

4 実験

本実験では, 対象タスクとして, IREX での定義での固有表現抽出 (**IREX**) と関根の拡張固有表現階層 Ver.7.1.0[7] での固有表現抽出 (**Extended_NE**) を用い, 訓練事例, 及び, Gold Standard として, それぞれ BCCWJ NE コーパス (2016 年 2 月 1 日版) と拡張固有表現タグ付きコーパスを用いた. また, 対象コーパスには, ClassA-1[8] に分類される 136 テキストを BCCWJ より抽出して用いた.

各タスクで使用した分類器やタグ付与モデルについては次のようになる. 分類器については, **IREX** では CRF++ Ver.0.58 (Linux 版)[9] を用いた. また, **Extended_NE** では, タグの種類が多いという理由から KyTea Ver.0.4.7 (Linux 版)[16][10] を用いた. パラメータは, **IREX** では KNP[17][11] の固有表現抽出モデルを作成する際のパラメータ[12]を, **Extended_NE** ではデフォルトのパラメータを用いた. また, KyTea のデフォルトモデルによる単語分割と同様に, EDR コーパス (日本語コーパス)[13] の基準に基づき, 用言の語幹と語尾を分割した. なお, **IREX** については, 比較用の分類器として KNP Ver.4.16 (Linux 版) を用いた. また, タグ付与モデルについては, 単語全体でひとつのモデルを生成する全体タグ付与モデルを用いた. また, ベースとなる素性は **IREX**, **Extended_NE** ともに, 表層, 品詞, 品詞細分類を用い, 品詞と品詞細分類については JUMAN++ Ver.1.01 (Linux 版)[18][14] による形態素解析結果を用いた.

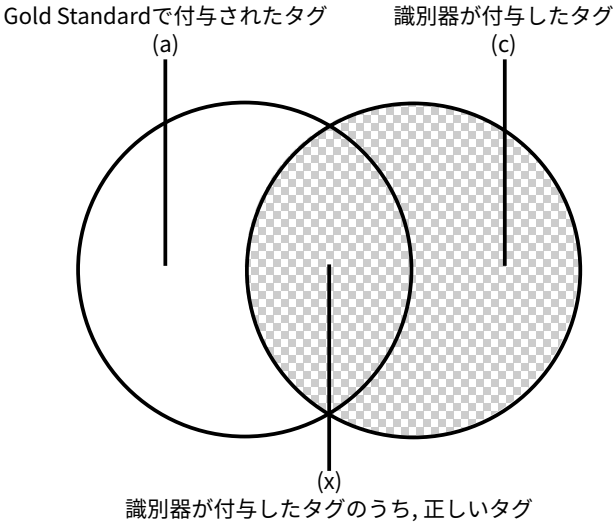

図 1: タグ集合

なお, 本実験では, できる限り多くのジャンルのテキストを含むような形で 5 分割交差検定を行っている. また, Gold Standard との比較による適合率 (精度), 再現率, F 値を指標として設定しており, タグ集合が図 1 のように示されるとき, それぞれ式 (1), 式 (2), 式 (3) のように与えられる.

$$p = \frac{n(x)}{n(c)} \quad (1)$$

$$r = \frac{n(x)}{n(a)} \quad (2)$$

$$f = \frac{2pr}{p+r} \quad (3)$$

5 結果

表 2, 表 3 は **IREX** を行った際の適合率, 再現率, F 値のマイクロ平均とマクロ平均を示している. **Extended_NE** → **IREX** は **Extended_NE** から **IREX** へ知識の転移を行った場合, **KNP** は KNP を用いた場合に該当する. また, **太字**は最も正解率が良い箇所を示しており, 下線部は **KNP** 以外で最も正解

[7] https://sites.google.com/site/extendednamedentityhierarchy/
[8] http://plata.ar.media.kyoto-u.ac.jp/mori/research/NLR/JDC/ClassA-1.list
[9] https://taku910.github.io/crfpp/
[10] http://www.phontron.com/kytea/index-ja.html
[11] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[12] http://www.lr.pi.titech.ac.jp/~sasano/knp/ne.html
[13] http://www2.nict.go.jp/3rd_Medium_Plan/out-promotion/techtransfer/EDR/J_index.html
[14] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++

表 2: 正解率のマイクロ平均 (**IREX**)

手法	適合率 (精度)	再現率	F 値
IREX	0.71	0.44	0.54
Extended_NE → IREX	**0.87**	0.66	**0.75**
KNP	0.78	**0.68**	0.73

表 3: 正解率のマクロ平均 (**IREX**)

手法	適合率 (精度)	再現率	F 値
IREX	0.45	0.30	0.36
Extended_NE → IREX	**0.54**	**0.41**	**0.46**
KNP	0.47	0.40	0.43

率が良い箇所を示している.

表 4, 表 5 は **Extended_NE** を行った際の適合率, 再現率, F 値のマイクロ平均とマクロ平均を示している. **IREX → Extended_NE** は **IREX** から **Extended_NE** へ知識の転移を行った場合に該当する.

6 考察

表 2, 表 3 より, **IREX** のマクロ平均と再現率以外のマイクロ平均について, **Extended_NE** の知識を転移させて学習する場合の正解率が **KNP** の正解率を上回っていることがわかる. KNP の訓練事例は本実験で用いた訓練事例よりも多いにも拘わらず, **KNP** の正解率を上回っている. また, 関根の拡張固有表現階層が IREX での定義を元に作られたという点から, IERX での定義のタグセットと関根の拡張固有表現階層のタグセットの間には強い関係性があると考えられる. そのため, **Extended_NE** のタグセットが **IREX** において, 性能を大幅に向上させるのに寄与していると考えられる.

さらに, 表 2, 表 3 より, **IREX** について, それ単体の知識のみで学習するよりも, **Extended_NE** の知識を転移させて学習する方が, マイクロ平均, マクロ平均ともに正解率が良くなっていることがわかる. しかし, 表 4, 表 5 より, **Extended_NE** について, **IREX** の知識を転移させて学習した場合の正解率は, それ単体の

表 4: 正解率のマイクロ平均 (**Extended_NE**)

手法	適合率 (精度)	再現率	F 値
Extended_NE	**0.64**	**0.17**	**0.27**
IREX → Extended_NE	**0.64**	**0.17**	**0.27**

表 5: 正解率のマクロ平均 (**Extended_NE**)

手法	適合率 (精度)	再現率	F 値
Extended_NE	**0.29**	**0.09**	**0.14**
IREX → Extended_NE	0.28	**0.09**	**0.14**

知識のみで学習した場合の正解率と同等以下の水準になっている. タグの種類という観点で見たとき, **IREX** よりも **Extended_NE** の方が種類が多いという点も加味すると, 強い関係性があるふたつのタスクについて, タグの種類が多いタスクから少ないタスクへの知識の転移では精度が向上するが, タグの種類が少ないタスクから多いタスクへの知識の転移では精度が向上しないということがわかる. これは, **IREX** でタグ A が付与されている固有表現に対し, **Extended_NE** ではより細かいタグ $A'_1, A'_2, ..., A'_n$ が付与されている状況によるものと推察できる. このような状況下では, **Extended_NE** でのタグが $A'_1, A'_2, ..., A'_n$ のいずれかであれば, **IREX** においては必ずタグ A となるが, その逆はどのタグになるか, 他の素性も見なければ推定できない. また, タグの数が多ければ, それぞれのタグごとのコーパス中の出現頻度が少なくなり, 訓練事例として寄与するためにはデータ量が必要となるが, 本実験の訓練事例サイズは比較的小さいため, あまり寄与できなかったと考えられる.

7 まとめと展望

本稿では, 素性に対して異なるタスクのタグセットを追加する方式のマルチタスク学習による学習手法 (タグセットの相互適応) の提案と既存手法との比較による考察を行った. 実験を通し, タグセットの間に強い関係性があるふたつのタスクについて, タグの種類が多いタスクから少ないタスクへの知識の転移では精度が向上するが, タグの種類が少ないタスクから多いタス

クへの知識の転移では精度が向上しないということがわかった．本稿では, 固有表現抽出のふたつのタスクで提案手法を適用したが, 固有表現抽出と語義曖昧性解消など, 種類の異なるタスクの間で提案手法を適用した場合についても, 今後比較していきたいと考えている．また, マルチタスク学習の先行研究にはニューラルネットワークを利用した研究があるため, 深層学習を利用した発展も考えている．

参考文献

[1] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, Vol. 22, No. 10, pp. 1345–1359, 2010.

[2] 神嶌敏弘．転移学習．人工知能学会誌, Vol. 25, No. 4, pp. 572–580, 2010.

[3] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

[4] Kanako Komiya, Daichi Edamura, Ryuta Tamura, Minoru Sasaki, Hiroyuki Shinnou, and Yoshiyuki Kotani. Domain Adaptation with Filtering for Named Entity Extraction of Japanese Anime-Related Words. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pp. 291–297, Hissar, Bulgaria, 2015. INCOMA Ltd. Shoumen, BULGARIA.

[5] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? *Advances in neural information processing systems*, pp. 640–646, 1996.

[6] Rich Caruana. Multitask learning. In *Learning to learn*, pp. 95–133. Springer, 1998.

[7] 鈴木雅也, 古宮嘉那子, 岩倉友哉, 佐々木稔, 新納浩幸. 固有表現抽出におけるアノテーション手法の比較. 研究報告自然言語処理 (NL), Vol. 2016, No. 7, pp. 1–8, 2016.

[8] Satoshi Sekine and Hitoshi Isahara. IREX: IR and IE Evaluation project in Japanese. In *Proceedings of the 2nd International Conference on Language Resources & Evaluation*, 2000.

[9] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended Named Entity Hierarchy. In *LREC*, 2002.

[10] Satoshi Sekine and Chikashi Nobata. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. In *LREC*, pp. 1977–1980, 2004.

[11] Satoshi Sekine. Extended Named Entity Ontology with Attribute Information. In *LREC*, pp. 52–57, 2008.

[12] Tomoya Iwakura, Ryuichi Tachibana, and Kanako Komiya. Constructing a Japanese Basic Named Entity Corpus of Various Genres. *ACL 2016*, pp. 41–46, 2016.

[13] Kikuo Maekawa. Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the 6th Workshop on Asian Language Resources (ALR)*, pp. 101–102, 2008.

[14] 橋本泰一, 乾孝司, 村上浩司. 拡張固有表現タグ付きコーパスの構築. 情報処理学会研究報告, 自然言語処理研究会報告 (NL-188-17), pp. 113–120, 2008.

[15] 橋本泰一, 中村俊一. 拡張固有表現タグ付きコーパスの構築-白書, 書籍, Yahoo! 知恵袋コアデータ．言語処理学会第 16 回年次大会発表論文集, 2010, pp. 916–919, 2010.

[16] Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pp. 529–533. Association for Computational Linguistics, 2011.

[17] Ryohei Sasano and Sadao Kurohashi. Japanese Named Entity Recognition Using Structural Natural Language Processing. In *IJCNLP*, pp. 607–612, 2008.

[18] 森田一, 黒橋禎夫．RNN 言語モデルを用いた日本語形態素解析の実用化．第 78 回全国大会講演論文集, Vol. 2016, No. 1, pp. 13–14, 2016.