意味役割付与におけるトランズダクティブ分野適応

大内 啓樹 [1,2]　鈴木 潤 [2,1]　乾 健太郎 [2,1]
[1] 理化学研究所 AIP センター　[2] 東北大学
hiroki.ouchi@riken.jp　{jun.suzuki,inui}@ecei.tohoku.ac.jp

1　はじめに

本稿では，解析したいテキスト (解析対象テキスト) 自体を学習に利用することによって，解析対象テキストに特化した意味役割付与 (Semantic Role Labeling; SRL) モデルを構築する．以下に概要を記す．

問題設定：　意味役割付与における分野適応．
提案手法：　単語分散表現のトランズダクティブ学習．
実験結果：　F1 値で約 1.5 ポイントの向上．
貢献 1：　解析対象テキストにモデルを直接的に適応させる初の試み．
貢献 2：　解析対象テキストを学習に用いるシンプルな手法によって，意味役割付与の性能向上を実証．

一般的な教師あり学習とは異なり，本研究で取り組むトランズダクティブ学習では，解析対象テキスト (テストデータ) を学習時に利用する．それにより，学習データと異なる分野のテキストを頑健に解析可能な意味役割付与モデルを構築可能であることを示す．

1.1　トランズダクティブ学習とは？

図 1 は本研究で取り組むトランズダクティブ学習の設定とその他の設定の比較を表している．一般的な教師あり学習の設定 (図 1 左) では，教師ラベル付き学習データでモデルを学習し，任意の未知テキストでその汎化性能を測定する．また，分野適応の設定 (図 1 中央) では，元分野と目標分野のデータからモデルを学習し，目標分野の未知テキストで汎化性能を測定する．どちらの設定においても，解析対象テキストは未知であると仮定される．一方でトランズダクティブ学習の設定 (図 1 右) では，解析対象テキストが所与であり，学習中にもそれらが利用可能な点である．解析対象テキストを学習時に効果的に利用できれば，解析対象テキストに対する予測が容易になると期待できる．

1.2　なぜトランズダクティブ学習に着目するのか？

一般的な教師あり学習などの設定では解析対象のテキストは未知であると仮定されてきたが，実応用上は必ずしもこの仮定を置く必要はない．例えば，現在多くの一般企業やユーザーが解析したいテキストデータを独自で保有しており，そのテキストデータ自体を高精度で解析したいというニーズがある．つまり，未知のあらゆるテキストの解析を目的とせず，手元にある特定のテキストのみを高精度で解析できれば良いという場合が少なくない．このように，解析結果の即時性は重要でなく，ある程度時間をかけたとしても (解析対象テキストが与えられてからモデル学習を始めても) 性能が最重要となる状況では，解析対象テキストを有効に利用した学習法が適していると考えられる．

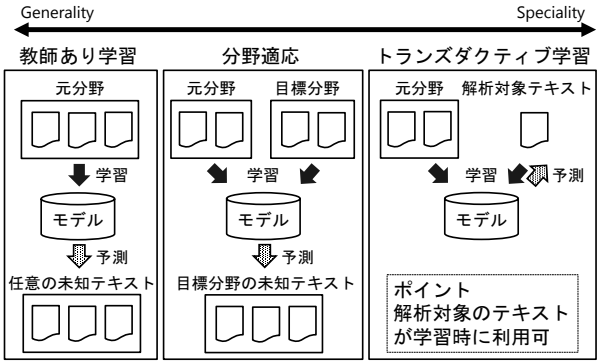

図 1　トランズダクティブ学習の問題設定の概略．

以上の理由から，従来研究のように未知のテキストにモデルを汎化させること (**generalization**) をめざすのではなく，本研究では解析対象の特定のテキスト集合にモデルを特化させること (**specialization**) によって，意味役割付与の精度向上をめざす．

2　関連研究

トランズダクティブ学習は 1990 年代後半に提唱され [4, 12, 7]，それ以降，自然言語処理分野のテキスト分類などにも応用されている [7, 11]．一方で，意味役割付与のような言語構造解析タスクに応用された例はほとんどなく，筆者の知る限り，本研究が初の試みである．関連研究として，意味役割付与における分野適応がある．これまでにいくつかの分野適応手法が提案されているが，限定的な性能向上にとどまっている [6, 13, 3, 5]．これらの研究は一般的な分野適応の設定であるため，本研究のトランズダクティブ分野適応の設定とは異なる．

3 問題設定

3.1 意味役割付与

入力文とターゲットの述語が与えられ，その述語の項を予測する．例として，次の文を考える．

$$\begin{array}{rcccc} \text{入力文：} & \text{She}_1 & \underline{\text{kept}_2} & \text{a}_3 & \text{cat}_4 \\ \text{正解の項：} & [\ \texttt{A0}\ ] & & [\quad \texttt{A1}\quad] & \end{array}$$

述語 kept の項として，She を動作主 (A0)，a cat を被動作主 (A1) として同定できれば正解となる．

本稿ではスパン選択問題として解く [8]．入力として，T 単語からなる文 $w_{1:T} = w_1, \cdots, w_T$ とターゲットの述語の位置 p が与えられる．出力として，ラベル付きスパンの集合 $Y = \{\langle i, j, r\rangle_k\}_{k=1}^{|Y|}$ を予測する．

$$\begin{aligned} &\textbf{入力}：X = \{w_{1:T}, p\} \\ &\textbf{出力}：Y = \{\langle i, j, r\rangle_k\}_{k=1}^{|Y|} \end{aligned}$$

各ラベル付きスパン $\langle i, j, r\rangle$ は，単語位置インデックス i と j，意味役割ラベル r から構成される．

上述した例文では，入力文は $w_{1:4}$ = She kept a cat であり，述語位置は $p = 2$ である．正解のラベル付きスパン集合は $Y = \{\langle 1, 1, \texttt{A0}\rangle, \langle 3, 4, \texttt{A1}\rangle\}$ である．この例文の可能なスパン集合は以下の通りである．

$$\begin{aligned} \mathcal{S}_{w_{1:4}} = \{&(1,1),(1,2),(1,3),(1,4),(2,2),\\ &(2,3),(2,4),(3,3),(3,4),(4,4)\}\ . \end{aligned}$$

これらの候補スパンの中から，各ラベルに対してスコアの最も高いスパンを選ぶ．

$$\underset{(i,j)\in\mathcal{S}}{\text{argmax}}\ \textsc{Score}_r(i,j),\ r \in \mathcal{R}\ .$$

ここで，関数 $\textsc{Score}_r(i,j)$ は，各スパン $(i,j) \in \mathcal{S}$ に対する実数値を返す．意味役割ラベル A0 として $(1,1)$ を選べれば正解である．同様に，ラベル A1 として $(3,4)$ を選べれば正解である．本稿では，最先端の SRL モデルである BiLSTM-span モデル [8] を関数 $\textsc{Score}_r(i,j)$ として用いる．

3.2 トランズダクティブ分野適応

一つの元分野から一つの目標分野への適応を考える．具体的には，以下のデータセットをモデルの学習に用いることを仮定する．

$$\begin{aligned} \mathcal{D}_{\text{src}}^{\text{train}} &= \{(X_i, Y_i)\}_{i=1}^{N_{\text{src}}^{\text{train}}}\ , \\ \mathcal{D}_{\text{tgt}}^{\text{unlab}} &= \{X_i\}_{i=1}^{N_{\text{tgt}}^{\text{unlab}}}\ , \\ \mathcal{D}_{\text{tgt}}^{\text{test}} &= \{X_i\}_{i=1}^{N_{\text{tgt}}^{\text{test}}}\ . \end{aligned}$$

$\mathcal{D}_{\text{src}}^{\text{train}}$ は元分野のラベル付き学習データであり，$\mathcal{D}_{\text{tgt}}^{\text{unlab}}$ は目標分野のラベルなしデータである．この二つのデータのみを利用して評価データのラベルを予測するのが「教師なし分野適応」である．これらに加え，目標分野の評価データ $\mathcal{D}_{\text{tgt}}^{\text{test}}$ も学習に利用可能である設定がトランズダクティブ分野適応である．注意点として，評価データの入力 (X) は学習の際に利用可能であるが，評価データの正解ラベル (Y) は利用できない．

4 提案手法

解析対象テキストの入力情報 (X) を有効に利用したい．入力情報を効果的にモデルに組み込む最も有効な方法の一つが単語分散表現の事前学習である．本稿では，解析対象テキストに特化した単語分散表現を学習する手法を提案する．

任意の単語分散表現 (SENNA[2] や GloVe[9] など) が使用可能であるが，本稿では最先端の意味役割付与モデル [8] で用いられている ELMo(Embeddings from Language Models)[*1][10] を用いる．ELMo は双方向 LSTM 言語モデルに基づいている．言語モデルの学習は対数尤度を最大化することによって行う．

$$\begin{aligned} \ell = \sum_{t=1}^{T} &\log p(w_t | w_{1:t-1}; \theta_x, \overrightarrow{\theta}_{\text{lstm}}, \theta_s) \\ &+ \log p(w_t | w_{t+1:T}; \theta_x, \overleftarrow{\theta}_{\text{lstm}}, \theta_s) \end{aligned} \quad (1)$$

ここで，θ_x は入力層のパラメータであり，$\overrightarrow{\theta}_{\text{lstm}}$ ($\overleftarrow{\theta}_{\text{lstm}}$) は LSTM のパラメータである．$\theta_s$ はソフトマックス層のパラメータである．

解析対象テキストが小規模である場合，言語モデルが上手く学習できないことが予想される．本提案手法では，データ量の不足を補いつつ，解析対象テキストに特化した言語モデルを学習する．具体的には，まず，大規模生コーパス $\mathcal{D}^{big}$ を用いて 1 式で言語モデルを学習する．次に，目標分野のテキスト $\mathcal{D}_{\text{tgt}}^{\text{test}}$ (または $\mathcal{D}_{\text{tgt}}^{\text{unlab}}$) を用いて同様に学習し，最終的な言語モデルを得る．得られた言語モデルをもとに各単語の分散表現を出力し，意味役割付与に用いる．

5 実験

5.1 データセット

CoNLL-2012 のデータセットを用いた．学習/開発/評価セットの分け方は既存研究と同様である．結果のスコア計算には CoNLL-2005 Shared Task で提供されている公式スクリプト[*2]を利用する．言語モデルの学習のための大規模生コーパスとして， One billion word benchmark[1] を用いる．言語モデルの学習は 10 エポック行う．意味役割モデルの学習は Ouchi ら [8] と同様に行う．

[*1] http://allennlp.org/elmo

[*2] http://www.lsi.upc.edu/∼srlconll/soft.html

	Baseline			(a)Unlab			(b)Test			(c)Test+Unlab		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BC	78.7	77.9	78.3	79.9	79.8	79.8	80.0	79.8	79.9	80.3	81.2	**80.8**
BN	79.0	81.4	80.2	78.2	81.8	79.9	78.2	80.9	79.6	79.3	81.1	**80.2**
MZ	78.8	79.2	79.0	78.6	79.9	79.2	79.5	80.5	80.0	80.2	80.4	**80.3**
PT	88.8	87.3	88.0	89.2	88.0	88.6	89.7	88.8	89.2	90.2	89.0	**89.6**
TC	75.3	74.4	74.8	74.7	77.7	76.2	75.2	75.7	75.5	75.9	78.0	**77.0**
WB	81.1	79.5	80.3	81.9	79.8	80.8	82.8	81.9	**82.3**	82.4	80.3	81.3
ALL	80.5	80.2	80.3	80.7	81.4	81.0	81.1	81.5	81.3	81.6	82.0	**81.8**

表 1　分野適応の結果 (NW → TGT)．元分野は「ニュース記事 (NW)」であり，目標分野は六つの分野のうちの各分野 (TGT=BC/BN/MZ/PT/TC/WB) である．

5.2　解析に用いる分野

元分野はニュース記事 (NW) とする．目標分野として以下の六つの分野のテキストを用いる．

- BC　Broadcast Convesation
- BN　Broadcast News
- MZ　Magazine
- PT　English Translation of the New Testament
- TC　Telephone Conversation
- WB　Weblogs and Newsgroups

5.3　比較手法

Baseline　：何も適応を行わないモデル．つまり，元分野のニュース記事 $\mathcal{D}_{\text{nw}}^{\text{train}}$ のみを学習に用いる．

(a) Unlab　：目標分野の生テキスト $\mathcal{D}_{\text{tgt}}^{\text{unlab}}$ に ELMo をフィットさせたモデル．

(b) Test　：解析対象テキスト $\mathcal{D}_{\text{tgt}}^{\text{test}}$ に ELMo をフィットさせたモデル．

(c) Test+Unlab　：解析対象テキスト $\mathcal{D}_{\text{tgt}}^{\text{test}}$ と目標分野の生テキスト $\mathcal{D}_{\text{tgt}}^{\text{unlab}}$ の両方に ELMo をフィットさせたモデル．

5.4　結果

表 1 はテストセットに対する分野適応の結果を示している．一列目は各分野を表し，二列目から最終列までが意味役割付与の適合率 (P)・再現率 (R)・F1 値 (F1) を表している．解析対象テキストに適応させた ELMo を用いたモデル (Test) が，適応させない ELMo を用いたモデル (Baseline) よりも高い F1 値を記録した．また，目標分野の生テキストを適応に用いたモデル (Unlab) も，適応しないモデル (Baseline) より高い性能を記録した．さらに，解析対象テキストと目標分野の生テキストの両方を用いて ELMo を適応させたモデル (Test+Unlab) は，解析対象テキストのみで ELMo を適応させるモデル (Test) よりも良い結果を記録した．これらの結果から，(i) 解析対象テキストに ELMo を適応させることの効果と，(ii) 目標分野の生テキストを追加的に用いることによる効果が確認できた．

	TGT		TGT + NW	
	F1	(c) との差	F1	(c) との差
BC	83.2	+2.4	85.0	+4.2
BN	83.1	+2.9	84.7	+4.5
MZ	80.5	+0.2	83.3	+3.0
PT	92.6	+3.0	93.4	+3.8
TC	83.0	+6.0	85.5	+8.5
WB	81.1	-0.2	84.2	+2.9

表 2　教師あり分野適応モデルの F1 値．

6　分析

6.1　目標分野の教師データを用いたモデルとの比較

本節では，教師あり分野適応の設定で構築したモデルと前節で構築したモデルの性能を比較する．教師あり分野適応では，目標分野の意味役割アノテーションを用いてモデルを学習する．このモデルの性能にどれほど近づけたかという指標は，教師なし分野適応における一つの目標 (上限) 値とみなすことができる．

表 2 は，目標分野の意味役割アノテーションを用いて学習したモデルの F1 値と，最も結果の良かったトランズダクティブ分野適応モデル (Test+Unlab) との F1 値の差 (diff) を示している．1 つ目の教師あり分野適応モデル (TGT) は，各目標分野の教師信号を用いて学習したモデルである．2 つ目のモデル (TGT + NW) は，各目標分野の教師信号に加え，「ニュース記事 (NW)」の教師信号を用いて学習したモデルである．

モデル TGT との比較: 1 つ目のモデル (TGT) とトランズダクティブ分野適応モデル (Test+Unlab) との性能差は，分野によって異なる結果となった．「雑誌 (MZ)」と「ウェブ記事 (WB)」では，ほとんど性能差がない．逆に，「電話会話 (TC)」では大きな差 (6 ポイント) が見られる．その他の分野に関しては，約 3 ポイント程度の性能差が見られる．以上の結果から，分野によっては，目標分野の教師信号を使って学習したモデルと同等の性能を持つモデルを提案手法によって構築できることがわかった．

	Baseline	Unlab	Test	Test+Unlab
BC	83.5	84.7	85.1	**85.6**
BN	85.5	85.4	85.2	**86.1**
MZ	85.6	86.1	86.6	**86.9**
PT	93.0	93.5	93.9	**94.0**
TC	79.8	81.4	80.1	**82.0**
WB	86.0	86.3	**87.6**	86.5
Avg.	85.6	86.2	86.4	**86.8**

表 3　スパン境界同定性能 (F1 値).

	Baseline	Unlab	Test	Test+Unlab
BC	93.8	94.3	93.9	**94.3**
BN	**93.8**	93.6	93.3	93.1
MZ	92.3	92.1	92.3	**92.4**
PT	94.6	94.7	95.0	**95.3**
TC	**94.3**	93.6	94.1	93.9
WB	93.4	93.7	94.0	**94.1**
Avg.	93.7	93.7	**93.8**	**93.8**

表 4　意味役割ラベルの正解率.

モデル TGT+NW との比較: 2 つ目のモデル (TGT+NW) とトランズダクティブ分野適応モデル (Test+Unlab) を比較する．これにより，ニュース記事分野に加えて目標分野の教師信号が利用可能である場合とそうではない場合の性能差を算出できる．結果として，「電話会話 (TC)」では 8 ポイント近い大きな性能差が見られたが，その他の分野では性能差を 5 ポイント以下に抑えることができている．したがって，分野によっては改善の余地が大きく残されていることがわかった．

6.2　スパン境界同定精度

トランズダクティブモデルが出力した結果に関して，スパン境界の一致のみを評価した F1 値を表 3 に示す．正解判定基準として，もし，予測されたラベル付きスパン $\langle \hat{i}, \hat{j}, * \rangle$ が，正解のスパン (i, j) と一致していればラベルに関わらず正解とみなす．結果として，ほぼすべての結果がベースラインの手法よりも改善している．特にモデル Test+Unlab は，その他のモデルよりも高い F1 値を示している．この結果から，ELMo による意味役割付与精度向上の要因として，スパン境界同定精度向上が大きく寄与していると考えられる．

6.3　意味役割ラベル予測精度

トランズダクティブモデルの出力結果に対して，意味役割ラベルのみを評価した正解率を表 4 に示す．評価対象として，スパン境界が正解と一致しているラベル付きスパンのみを扱う．結果として，トランズダクティブモデルとベースラインとの間に大きな性能差は見られなかった．いずれの分野においても，各モデル間の性能差は 1 ポイント以内にとどまっている．興味深いことに，スパン境界同定精度 (6.2 節) において性能差が見られた結果とは対照的な結果となった．

一つの理由として，SRL タスクの持つ統語的性質と意味的性質の違いがあると考えられる．スパン境界は，句とほぼ一致するため，統語的な側面と関係が深い．一方，意味役割ラベルはタスクの意味的性質をより反映している．例えば，統語的には「主語」であるスパンに異なる意味役割ラベルが付与されることも少なくない．つまり，タスクの定める意味役割に依拠する比重が大きくなる．これら両側面が意味役割付与には求められ，最終的な結果に影響を与える．そのうちの統語的側面の精度向上に，本トランズダクティブ学習手法は特に寄与していると考えられる．

7　おわりに

本稿では，解析対象テキストに特化した単語分散表現の学習法を提案し，意味役割付与モデルの性能向上に取り組んだ．それにより，学習データと異なる分野のテキストを頑健に解析可能な意味役割付与モデルを構築可能であることを実験的に示した．今後の課題として，目標分野における少量の教師信号が利用可能である設定での性能評価が挙げられる．

参考文献

[1] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

[2] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. In *Journal of Machine Learning Research*, 2011.

[3] Quynh Thi Ngoc Do, Steven Bethard, and Marie-Francine Moens. Domain adaptation in semantic role labeling using a neural language model and linguistic resources. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(11):1812–1823, 2015.

[4] Alexander Gammerman, Volodya Vovk, and Vladimir Vapnik. Learning by transduction. In *Proceedings of UAI*, pages 148–155, 1998.

[5] Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych. Out-of-domain framenet semantic role labeling. In *Proceedings of EACL*, pages 471–482, 2017.

[6] Fei Huang and Alexander Yates. Open-domain semantic role labeling by modeling word spans. In *Proceedings of ACL*, pages 968–978, 2010.

[7] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of ICML*, pages 200–299, 1999.

[8] Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. A span selection model for semantic role labeling. In *Proceedings of EMNLP*, pages 1630–1642, 2018.

[9] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543, 2014.

[10] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[11] Hirotoshi Taira and Masahiko Haruno. Text categorization using transductive boosting. In *Proceedings of ECML*, pages 454–465, 2001.

[12] Vladimir Naumovich Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[13] Haitong Yang, Tao Zhuang, and Chengqing Zong. Domain adaptation for syntactic and semantic dependency parsing using deep belief networks. *Transactions of ACL*, 3:271–282, 2015.