Corpus Sharing: 異なるアノテーション体系への自動変換手法

荒牧英治†
†東京大学 知の構造化センター
eiji.aramaki@gmail.com

1 はじめに

アノテーション・コーパスは言語処理の基礎的リソースであり，毎年多くのコーパスがリリースされている．それらのいくつかは非常に類似した目的をもつが，アノテーション・カテゴリー（以下カテゴリ）が異なるため，それらを併せて用いることができないことが多い．

例えば，固有表現抽出では MUC スタイルのアノテーション（7 種類のカテゴリ）を採用したコーパスが多いが，IREX では＜人工物＞が扱われ，カルテ・コーパス [6] では，＜人名＞が＜患者名＞と＜医師名＞に細分化されている．また，Sekine 等は百以上の階層化した固有表現カテゴリを提案している [5]. このようなカテゴリーの差異がある場合，複数のコーパスを併せて使用することは困難である．

この問題に対応するため，本稿では，異なるカテゴリを持つ２つのコーパスを相互に変換する手法を提案する．提案手法は次の２つのステップからなる．

STEP1: まず，両コーパスでそれぞれ学習器のトレーニングを行う．次に，学習器はそれぞれ相手側のコーパスでラベルを予測する．予測されたラベルはオリジナルのラベルと比較され差異があれば，その箇所に新ラベルを与える．

例えば，図 1 左のような２つのコーパスが与えられたとする (Corpus-A は “○” と “×” のラベル ; Corpus-B は “○” と “+” のラベルを持つ)．ラベルの差異のため，ラベルの衝突 ((A:B)=(○:+) または (×:+)) が起き，これらに対して，新ラベルを付与する (例えば (A:B)=(○:+) に “△”; (A:B)=(×:+) に “∗”).

この際，オリジナルのラベルを復元できるように “△” は Corpus-A の “○” に，Corpus-B の “+” に対応するといったマッピングを作成しておく．

STEP2: STEP1 で得られた新ラベルを持つコーパスを用いて，再び，学習器をトレーニングし，相手側コーパスのラベルを推定する．もし，推定が誤り，かつオリジナルのラベルが復元できる場合はコーパスのラベルを書き換える．

例えば，Corpus-B の “∗” を Corpus-A 由来の学習器が “△” だとミスプレディクトしたとする．△はそもそも Corpus-B では ∗ を意味していたため，この両ラベルの違いはオリジナルラベルへの復元を損ねない．この場合，Corpus-B の “∗” を “△” と書き換える．

上記の処理で，提案手法は両コーパスに対して元のラベルに変換可能な新ラベルを付与し，併用を可能とする．

2 提案手法

表記の簡便のため，本稿ではコーパスのインスタンス列を $X = (x_1...x_n)$，インスタンスのラベル列を $Y = (y_1...y_n)$，コーパスを $(X : Y) = (x_1 : y_1)...(x_n : y_n)$ と記述する．学習器は $f(x_i) : x_i \mapsto y_i$ を学習する．

解くべきタスクは，２つのコーパス (Corpus-A $(X_a : Y_a) = (x_{a1} : y_{a1})...(x_{an} : y_{an})$ と Corpus-B $(X_b : Y_b) = (x_{b1} : y_{b1})...(x_{bm} : y_{bm})$) が与えられた際に，ラベルの互換性がもっとも高い新しいラベル列 (Y'_a と Y'_b) を両コーパスに付与することである．ラベルの互換性は，未知のインスタンス X に対しての一致率 (i.e., $f_a(X) = f_b(X)$) にて観測することができる．

このラベル付与には次のような制約がある．新しいラベル (Y'_a と Y'_b) はオリジナルのラベル (Y_a と Y_b) に復元できなくてはならない．これはオリジナルのラベルはそれぞれのコーパスのポリシーにて作成されているため，保存される必要があるからである．

提案手法は次の２つのステップからなる．

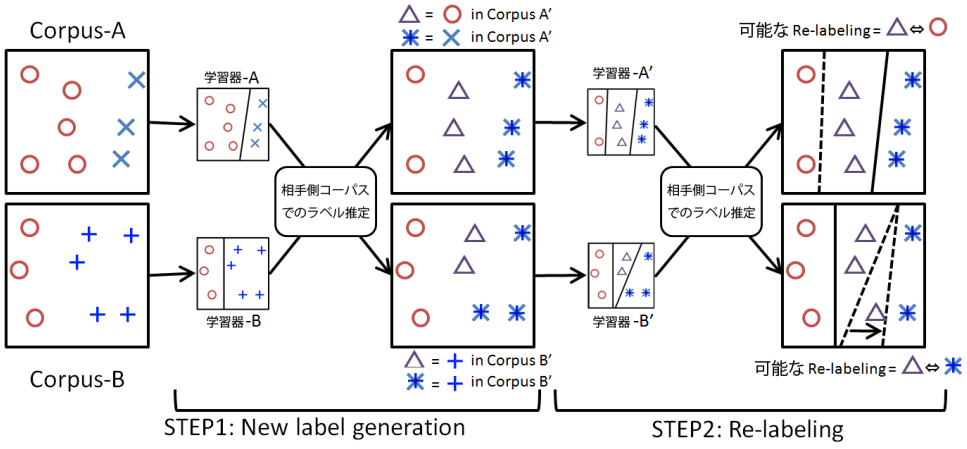

図 1: 提案手法の流れ.

STEP 1: 新ラベルの生成

まず，2つの学習器をそれぞれのコーパスでトレーニングし $f_a : X_a \mapsto Y_a$ と $f_b : X_b \mapsto Y_b$ を得る．

次に，それぞれの学習器が他方のコーパスでラベルを推定する．もし，推定誤り（オリジナルのラベルと推定されたラベルが異なる）が起こると (i.e., $F_a(x_{bi}) \neq y_{bi}$ または $F_b(x_{ai}) \neq y_{ai}$), その異なりを新ラベル候補とする．

最後に，新ラベル候補を用いて，両コーパスで学習器を訓練しなおし，未知なデータ（テストセット）に対して，両学習器がラベル推定を行う．もし，両学習器の推測したラベルの一致率が低いならば，そのラベルは学習困難であり，不適切なラベル（ノイズ）の可能性が高い．そこで一致率が閾値（後述する実験では40%と設定した）よりも高いラベルを採用し，採用されなかった新ラベル候補は，採用されたラベルセットのうち最も確率の高いラベルに置換した*.

STEP 2: ラベル修正

より正確にラベルの性質を一致させるために，本ステップではラベルの修正を行う．まず，再び新しい学習器をそれぞれのコーパスでトレーニングする ($f'_a : X_a \mapsto Y'_a$ and $f'_b : X_b \mapsto Y'_b$). 次に，Step 1 と同様に相手側コーパスのラベルを推定する．もし，推定されたラベル $f'_b(x_{ai})$ とコーパスのラベル y'_{ai} が異なりかつ，両者（$f'_b(x_{ai})$ と y'_{ai}）が同じラベルにマッピングされているならば，コーパスラベルを推定されたラベル $f'_b(x_{ai})$ に修正する．

*この置換を効率よく行うために，あらかじめ n-best 解を推定しておく．

表 1: 学習器の素性.

語彙
POS
suffix (1..2 文字)
prefix (1..2 文字)
case (大文字，小文字，両方)
length (# of characters)
時間／日付表現に関するアドホックな素性

* POS が付与されていないコーパスに対しては，タガー [7] により POS を付与した．

3 実験

2 種類の実験（疑似的なコーパスを用いた実験と実際のコーパスを用いた実験）を行った．

3.1 疑似コーパスでの実験

【実験設定】CoNLL02 トレーニングセット（2カ国語; NED と ESP) †を 2 分割し，図 2 の処理にて疑似的にアノテーションが異なるコーパスを生成した．

トレーニングにあたっては前後 2 語の表 1 の素性を用い，CRF++‡ を用いて学習を行った．

【比較手法と評価】次の 3 つの手法を比較した．

1. **ORG**: オリジナルのコーパスのみ，

†http://www.cnts.ua.ac.be/conll2002/ner/
‡http://www.chasen.org/ taku/software/CRF++/

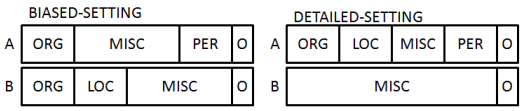

図 2: 擬似的に生成されたコーパス.

BIASED 設定: Corpus-A のすべての LOC を MISC に置換し，Corpus-B のすべての PER を MISC に置換する.

DETAILED 設定: Corpus-B のすべての ORG, LOC と PER を MISC に置換し，Corpus-A はそのまま用いる.

表 4: 各コーパスのアノテーション

MUC	CONLL	I2B2
ORGANIZATION	ORG	HOSPITAL
LOCATION	LOC	LOCATION
PERSON	PER	DOCTOR
TIME	MISC	PATIENT
DATE		DATE
MONEY		ID
PERCENT		
153,660 語	219,553 語	220,209 語

2. **TRANS**: STEP1 にて新ラベルをふった両方（corpusA および B）のコーパスを用いる，

3. **TRANS+**: STEP2 にて新ラベルをふった両方のコーパスを用いる（提案手法）.

評価にあたっては，Corpus-A または B のラベルセットに変換し，各コーパスのテストセットを用いて行った（したがって，2 通りの精度が得られる）.

【結果】 実験の結果（表 2 と表 3），多くの指標で TRANS+がもっとも精度が高い．このことは，理想的な状態では提案手法で相手側のコーパスを用いる利得が，ラベル数の増加や推定誤り（ノイズ）による損失を上回ることを示している.

生成されたラベルは，BIASED 設定では *-MISC-vs-B-LOC や*-PER-vs-*-MISC が生成され，DETAILED 設定では*-ORG-vs-*-MISC ，*-LOC-vs-*-MISC や*-PER-vs-*-MISC が生成された．擬似的に置換されたラベルを元に復元できていることがわかる.

3.2 実際のコーパスでの実験

【実験設定】次の 3 種類のコーパスでの実験を行った: (1)**MUC**: MUC6 と MUC7 の NE コーパス[§], (2)**CONLL**: CoNLL2003 の英語 NER コーパス[¶], (3)**I2B2**: カルテ文章の個人情報匿名化コーパス[‖]．各コーパスにアノテーションされたカテゴリを表 4 に示す.

学習器は Averaged Perceptron [1] (AP), Passive Agressive [2] (PA), Confidence Weighted [3] (CW) の 3 手法[**] を比較した．トレーニングにあたって，前節と同じ素性にてタグ毎に one-vs-rest 法を用いた．実験では ORG と TRANS+を F 値で評価した.

【結果】 結果を表 5(AP), 表 6(PA) と表 7(CW) に示す．提案手法により，CONLL+MUC (CW, MUC スタイル側の評価) が精度向上見せたが，他の場合は精度が下がる傾向があった.

例えば，MUC コーパスとの組み合わせ，AP, PA では提案手法は大きく精度を下げた．これらの設定では，そもそもの推定精度が低い（例えば AP-ORG や PA-ORG）場合が多い．この事実から，提案手法が働くためには，ある程度の推定精度が必要であると考えられる.

また，I2B2 との組み合わせも精度を得られなかった．この原因は，I2B2 コーパスと他のコーパスの間にある次の差異が原因だと思われる：

1. **カテゴリの差異**: カルテ独自のカテゴリ（医師名 DOCTOR，患者名 PATIENT など）が新聞に出現する頻度は少ない.

2. **スタイルの差異**: カルテで頻出する日付表現 DATE スタイル（「11/23/04」など）が新聞で出現する頻度は少ない．逆に，新聞で出現する現時点からの相対的な日付表現（「Last Saturday」など）がカルテに出現する頻度は少ない.

3.3 考察

3.1 節の疑似コーパスを用いた実験では，理想的な状態では，提案手法は高い精度でラベル変換が行え，相手側コーパスを用いて対象コーパスの精度を向上できることが分かった.

3.2 節の実際のコーパスを用いた実験では，提案手法の 2 つの限界が示唆された．(1) 十分な精度でラベルを推定できない場合，提案手法は精度を得られない．(2) 対象となるコーパスのカテゴリ／スタイルが大きく異なる場合は，提案手法は精度を得られない.

[§] http://www-nlpir.nist.gov/related_projects/muc/

[¶] http://www.cnts.ua.ac.be/conll2003/

[‖] https://www.i2b2.org/NLP/

[**] http://code.google.com/p/oll/

表 2: BIASED 設定結果.

		ORG	TRANS	TRANS+
NED	P	0.757	0.762+++	**0.765**+
(Corpus-A)	R	0.728	0.732	**0.737**
	F	0.742	0.747	**0.751**
NED	P	0.759	0.771+++	**0.774**
(Corpus-B)	R	0.734	0.741	**0.744**
	F	0.747	0.756	**0.758**
ESP	P	0.748	0.764+++	**0.773**+++
(Corpus-A)	R	0.742	0.757	**0.765**
	F	0.745	0.760	**0.769**
ESP	P	0.743	0.747+++	**0.749**
(Corpus-B)	R	0.733	0.740	**0.743**
	F	0.738	0.743	**0.746**

表 3: DETAILED 設定結果.

		ORG	TRANS	TRANS+
NED	P	0.727	0.726	**0.730**+++
(Corpus-A)	R	0.698	0.703	**0.707**
	F	0.712	0.714	**0.718**
NED	P	0.909	0.917+++	**0.920**+++
(Corpus-B)	R	0.882	0.887	**0.891**
	F	0.895	0.901	**0.905**
ESP	P	0.742	0.746+++	**0.749**
(Corpus-A)	R	0.732	0.744	**0.747**
	F	0.737	0.745	**0.748**
ESP	P	0.908	0.918+++	**0.920**
(Corpus-B)	R	0.909	0.913	**0.915**
	F	0.907	0.915	**0.918**

Corpus-A と Corpus-B は用いたテストセットを示す（それぞれタグのカテゴリが異なる）．+ は McNemar 検定 [4] での統計的有意を示す（TRANS は ORG と比較；TRANS+は TRANS と比較；+++:$p < 0.01$, ++:$p < 0.05$, +:$p < 0.1$ ）（適合率でのみ調査した）．太字は最高の値を示す.

表 5: AP の結果.

CONLL	63.61	+MUC	63.41
		+I2B2	63.10
MUC	56.34	+CONLL	55.42
		+I2B2	56.46
I2B2	74.30	+MUC	74.00
		+CONLL	66.82

表 6: PA の結果.

CONLL	61.89	+MUC	60.02
		+I2B2	56.24
MUC	55.25	+CONLL	55.36
		+I2B2	52.85
I2B2	67.73	+MUC	50.17
		+CONLL	**68.45**

表 7: CW の結果.

CONLL	80.50	+MUC	**81.58**
		+I2B2	80.57
MUC	66.96	+CONLL	66.17
		+I2B2	64.27
I2B2	81.87	+MUC	81.88
		+CONLL	80.50

上記のような制限はあるものの提案手法はコーパスラベルを変換する差異のツールとしては有用であろうと考えている．

3.4 今後の課題

今後の課題としては次の 3 つが考えられる．まず，本研究は提案手法の理論的裏付け／数学的定式化を欠いており，新ラベル生成のための閾値 (TH) などアドホックに設定している．これを補足することが第一の課題である．

次に，本研究では NE タスクを扱ったが，他のラベル付きコーパスを用いたタスク（文章分類や依存構造解析）などにおいても提案手法が適応可能かどうかの調査も必要である．

最後に，提案手法は複数回のトレーニング／テストの繰り返しが必要で，計算コストが大きい．今後の高速化が求められる．

4 おわりに

本稿は，異なるラベルをもつコーパスを同じコーパスに変換する手法を提案した．手法はあらゆるラベル付きコーパスに適応可能な汎用的なものである．実験では，理想的な状態で提案手法が有効であること，また，有効でない場合の条件を実証的に示した．本手法を用いることで，今後，コーパス変換／併用の効率が上がることを期待している．

参考文献

[1] Michael Collins. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proceedings of the Conference on Empirical Methods in Naturl Language Processing (EMNLP2002)*, pp. 1–8, 2002.

[2] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *The Journal of Machine Learning Research*, Vol. 7, pp. 551–585, 2006.

[3] Mark Dredze, Koby Crammer, and Fernando Pereira. Confidence-weighted linear classification. In *Proceedings of the International Conference on Machine Learning (ICML2008)*, pp. 1–8, 2008.

[4] L. Gillick and SJ Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 532–535, 1989.

[5] Satoshi Sekine, K. Sudo, and C. Nobata. Extended named entity hierarchy, 2002.

[6] Sibanda Tawanda and Uzuner Ozlem. Role of local context in automatic deidentification of ungrammatical, fragmented text. In *Proceedings of the Human Language Technology conference and the North American chapter of the Association for Computational Linguistics (HLT-NAACL2006)*, pp. 65–73, 2006.

[7] Yoshimasa Tsuruoka and Jun'ichi Tsujii. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of HLT/EMNLP*, pp. 467–474, 2005.