共起性を考慮した素性集合分割による Co-training
— 固有表現抽出の場合 —

辻克俊 † 権瓶竹男 ‡ 森辰則 ††
† 横浜国立大学 大学院 工学研究科 ‡ 横浜国立大学 工学部 †† 横浜国立大学 大学院 環境情報研究院
E-mail: {katsu,gompei,mori}@forest.eis.ynu.ac.jp

1 はじめに

固有表現抽出規則を学習する場合、学習用に分類に関する正解が付いた訓練データを用いて機械学習を行なう手法が提案されている。以下では分類に関する正解をラベルと呼ぶことにする。 しかし、そのデータ作成は、膨大な人的労力を必要とする。そこで、少量のラベル付き訓練データを用いて、大量のラベルなしデータを分類し、ラベル付き訓練データを擬似的に増やすことが検討されている。 Co-training[5] はその様な学習手法の一つである。 本稿では、 Support Vector Machine[1] に注目し、これを Co-training で用いる方法について述べる。そして、分類器を学習する前に素性集合を分ける方式として、素性間の共起性に基づく方法を含む幾つかの手法を検討する。

2 日本語固有表現抽出手法

本稿では、 IREX 日本語固有表現抽出タスクで使用されている 8 つの固有表現 (人名、地名、組織名、固有物名、日付、時間、割合、金額) について、以下に述べる方法により抽出を行うものとする。

2.1 Chunking の設定

本稿では固有表現抽出タスクを、形態素解析された形態素列に対して固有表現に関するラベルを振っていくタスクであると考える。複数の形態素の列により一つの固有表現が構成されることがあるので、そのラベルにはチャンク化 (chanking) の情報を持たせるのが普通である。本稿では山田ら [1] の IOB2 方式を採用している。

2.2 SVM による固有表現抽出

ラベル付き訓練データから固有表現抽出規則を学習する際に、 Support Vector Machines(SVMs) という機械学習アルゴリズムを採用する [1]。

2.2.1 素性、事例ベクトルの設定と解析方向

各事例は文中の一つの形態素に対応する事例ベクトルとその形態素に付与されるべきラベルの組で表現される。事例ベクトルは当該形態素を中心とした一定の窓幅の形態素に関する情報から構成される。例えば後の実験では窓幅を 5 にしているが、この場合は、図 1のように当該形態素の前後 2 形態素を考慮することになる。 また、文の先頭より一つ前に「文頭」、文の末尾より一つ後に「文末」という仮想の形態素が存在するとしている。事例ベクトルの素性は、窓内の各形態素について、その表層表現、品詞、構成する文字種の各々をその形態素の現

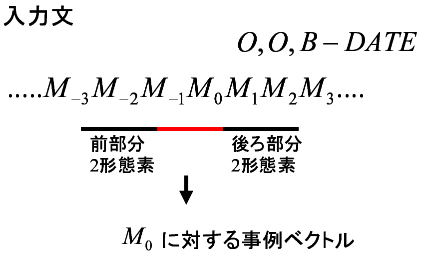

図 1: 窓幅 5 の左向き解析

れる位置と組み合わせたものを用いる。後の実験では、素性に採用する表層表現は、訓練に用いるコーパスにおいて 5 以上の頻度のものを採用し、構成された素性について、 3 以上の頻度のもの事例ベクトルの作成に使用した。

さて、固有表現の推定を文頭から進めていく場合、ある特定の形態素についてのラベル推定において、直前までの形態素にはすでに推定されたラベルが付与されているので、これを用いることができる。この素性は当該形態素のラベル推定に非常に効力を発揮することが報告されている。この解析方式を「右向き解析」と呼ぶ [1]。同様にして、文末から解析を進める場合には、「左向き解析」と呼ぶ。以下の実験では、高精度であることが報告されている左向き解析を採用する。

2.2.2 多値分類問題の適用法

SVM は本質的に二値分類を行なう分類器を生成するが、これを多値分類に拡張する手法として、 pairwise 法、one v.s. rest 法が提案されている [2]。 以下に述べる Co-training では、未知事例を分類した時に信頼度の高い事例を新たにラベル付き事例として採用するので、分類時の信頼度の評価が重要である。 pairwise 法は、信頼度が投票数という離散値であり、分解能が低いために Co-training には適さない。一方、 one v.s. rest 法は分離平面からの距離を信頼度とするので信頼度の分解能が高く、 Co-training に適していると考えられる。よって、本稿は、 one v.s. rest 法を用いる。

3 SVM と Co-training に基づく固有表現抽出

3.1 Co-training

Co-training は、独立な 2 つの素性集合を設定し、各々の素性集合のみを用いて、ラベル付訓練データからそれぞれ分類器 A、 B を作成する。そして、各々の分類器を用いて、ラベルなし訓練データの判別をそれぞれ行っ

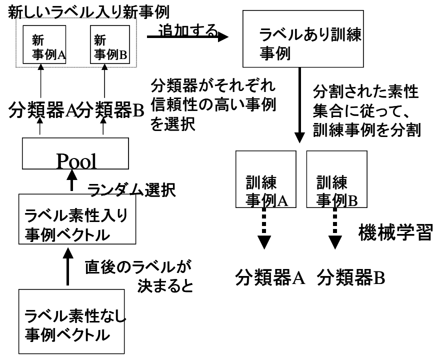

図 2: SVM と Co-training に基づく固有表現抽出

て、信頼性の高いものをラベル付き訓練データに加えてゆくことで分類器の精度を向上させてゆく。

3.2 Co-training における素性集合間の独立性

Collins ら [4] は英語の固有表現抽出について、対象となる単語から得られる素性集合とその前後の文脈情報からえられる素性集合に分ける方法を提案している。英語では、単語の頭文字が大文字であるか否かが固有表現判定に重要な手がかりとなるので、当該単語の素性だけでもある程度の学習が可能であると考えられる。一方で、日本語の場合、そのような慣習はないので当該単語の情報だけでは分類が十分に進まない可能性がある。

新納 [3] は語義判別タスクにおいて前文脈と後文脈に分割、これら素性集合の間での独立性を各事例について検査し、独立性の低い事例については採用しないようにしている。この方法は、独立性の保証はなされるものの、特定の事例を捨て去るために学習できない文脈が存在する可能性が高い。

そこで我々は特定の事例を捨てるのではなく、事前に素性集合を適切に分割する方法を検討する。

3.3 アルゴリズム

SVM と Co-training に基づく固有表現抽出について、その概略を図 2に，そのアルゴリズムを以下に示す。このアルゴリズムは，ラベルつき事例の集合から出発し，ラベルなし事例集合に対して，すこしずつ確からしいと思われるラベルを振っていく．これにより，最初のラベルつき事例よりも多いラベルつき事例を得る．これを通常の SVM に基づく固有表現抽出システムの学習事例とすることにより，最初の事例集合の場合よりも精度の高い分類器が構成できることが期待される．

1. 少量のラベル付き訓練事例集合 L と大量のラベルなし訓練事例集合 U を用意する。
2. ラベルなし訓練事例集合 U から一定数の事例をランダムサンプリングし、これを事例プール P とする。
3. ラベル付き訓練事例集合における素性の頻度分布を基に、素性集合間の相互依存性が低くなるように、2 つの素性集合 X_1、X_2 に分割する。
4. L より X_1 中の素性のみを持つラベル付き訓練事例集合 L_1 と，X_2 中の素性のみを持つラベル付き訓練事例集合 L_2 を生成する。
5. 分割されたラベル付き訓練事例集合 L_1、L_2 について，各々，SVM を用いてそれぞれ分類器 h_1,h_2 を学習する。
6. h_1 を使って P の事例にラベル付けを行ない、信頼性の高いものから順に，ラベル付き訓練事例集合 L のラベルの頻度分布に従って、n 個選ぶ。
7. h_2 を使って P の事例にラベル付けを行ない、信頼性の高いものから順に，ラベル付き訓練事例集合 L のラベルの頻度分布に従って、p 個選ぶ。
8. 6 と 7 で得られたラベル付の事例 $n+p$ 個を事例プール P から取り除き L に加える。
9. 事例プール P に、ラベルなし訓練事例集合 U から無作為抽出した $n+p$ 個の事例を追加する。
10. 3 から繰り返す。

このアルゴリズムにおいて、事例プール P の大きさ、ならびに、分類された未知事例のうち、上位何位までを信頼して加えるかによって、新規追加事例の質が変わってくる。 また我々の場合には左向き解析を行なうので、後に述べるように未知事例の扱いに工夫が必要である。

3.4 素性集合間の従属性を考慮した素性分割

先に述べたように Co-training が有効に働くためには、二つの分類器で利用される素性集合間の独立性が高い (従属性が低い) ことが必要である。 2 組の素性集合同士の独立性が低くなると、一方の分類器の採用した新規事例が他方の分類器に対して新しい素性の組み合わせをもたらさず、新たな事例の追加がない状況に陥り、各分類器の精度が頭打ちになる。

1. 素性集合間の共起性を計り、これが低くなるように素性集合を分割する方法。

   この手法は、素性集合間の独立性が高くなることを保証できるが、その一方で、各素性集合内での従属性が強くなるようになるので、各文脈毎に見た時に、素性がどちらか一方の素性集合に偏る可能性がある。

2. 素性の頻度分布が両素性集合で均一化するようにランダムに振り分けて配置す る方法。

   この手法は、積極的に素性間の独立性を保証する分け方をするのではなく、素性分布を均一化することにより、素性集合内と素性集合間の従属性 (独立性) がほぼ等しくなるようにするものである。

前者については、素性集合間の従属性を表す尺度が必要となるが、本稿では、素性間の共起性に基づく次の尺

度を使用する。二つの素性集合 X_1 ならびに X_2 の間に存在する共起性 $Cor(X_1, X_2)$ を以下のように定義する。

$$Cor(X_1, X_2) = \sum_{i=1}^{m}\sum_{j=1}^{n} dice(l_i, r_j)$$

m: 素性集合 X_1 の要素数
n: 素性集合 X_2 の要素数
l_i: 素性集合 X_1 の要素
r_j: 素性集合 X_2 の要素

また，$dice(l_i, r_j)$ は以下のように定義される dice 係数である。

$$dice(l_i, r_j) = \frac{2frq(l_i, r_j)}{frq(l_i) + frq(r_j)}$$

$frq(x)$: L の中で現れた x の頻度
$frq(x, y)$: L の中で x と y が同じ事例の中に共起した頻度

しかし、素性集合間の独立性だけに注意をすると、分割した 2 つの素性集合において、ある事例ベクトル a が与えられた時に、事例ベクトル a の情報を持つ素性がすべてどちらか片方に偏ってしまう可能性がある。この問題に対して、我々は、素性の種類が多い表層表現に関する素性とそれ以外の種類の少ない素性について異なる扱いをすることにより、緩和を試みる。

種類が少ない、品詞 - 位置の組ならびに文字種 - 位置の組に関する素性は、同じ位置の素性を一つのグループとして扱い、二つの素性集合に分割するときにはこれを不可分な単位として考える。二つの素性集合に分ける時には、前半二形態素と当該形態素に対応する組と、後半二形態素に対応する組に、それぞれ、分けている。

一方、種類が非常に多い、表層表現 - 位置の組に関する素性については、位置によるグループ化を行なうと非常に大きな素性集合となってしまい、素性集合の設定において制御が難しい。そこで、各グループを素性の出現頻度が均等になるように n 個に分割し、これを単位として素性集合の独立性が高くなるように割り振る。

共起性の尺度に基づく分割の場合には、まず、初期状態として、各素性集合の要素数が同じ程度になるように、表層表現 - 位置の組に関する素性のグループをランダムに振り分ける。次に、山登り法を用いて、集合間の独立性が高くなる組み合わせを探索する。

3.5 ラベル推定に対する信頼度

Co-training においては，ラベルなし事例の分類結果において，ラベル付与の信頼度の高い事例をラベル付きとみなして採用していく．この時に，どのようにラベル付与の信頼度を与えるかが問題となる．

分類器がラベルなし事例にラベル付けを行う際に、one v.s. rest 法では、複数の候補のうち、二値分類器の決定関数の値が最大となる分類器に対応するラベルを採用する [2]。よって、本手法でも、各事例の分類結果としては、決定関数の値の最大値をとるラベルを採用するが、この時の決定関数の値をそのラベルに対する信頼度として扱う。

3.6 解析方向を考慮した事例の選択手法

左向き解析では、事例ベクトルの中に、直後ならびにその先の形態素に対するラベルの推定結果が入っているために、それら形態素に対するラベルの推定が終らない限り、分類することができない。Co-training においては、これを考慮して、ラベルなし訓練データの集合 U から直接、事例プール P への追加を行なうのではなく、U 中の事例を処理の進度にあわせて、利用可能なラベルなし訓練データの集合 U' へ移動し、そこから事例プール P への追加を行なう。先に示したアルゴリズムの変更点は以下の通りである。

1. 初期の事例プールの作成
   (a) ラベルなし訓練事例集合 U から、文の最後の形態素に対応する事例 (これは既に利用可能) を選択し、利用可能なラベルなし訓練事例集合 U' とする。
   (b) U' から一定数の事例をランダムサンプリングし事例プール P とする。
2. 新たに採用された事例による利用可能なラベルなし訓練事例集合 U' の更新

   新たに高信頼度の事例を L に加える場合には、その事例の直前の形態素に対応する事例について、直後の推定ラベルを表す素性を設定することができるので、これを埋め、U から U' に移動する。
3. プールの更新

   プールの更新においてはその不足分を U からではなく、U' から無作為に抽出する。

4 実験および考察

4.1 実験環境と方法

実験では、独立行政法人通信総合研究所で作成された CRL 固有表現データを使用した。また、形態素解析器としては ChaSen Ver.2.02 を使用した。

CRL 固有表現データからは全部で 26880 個のラベル付き事例が得られる。初期学習の精度の影響を調べるために，本実験ではこの事例集合から、初期学習のためのラベル付き事例集合 L として 10000 個を選んだ場合と，20000 個を選んだ場合の 2 種類を調べる．学習結果の評価用にこれとは別の 10000 個の事例を用意した。Co-training に用いるラベルなし事例集合 U には、これらとは異なる 55000 個を利用した。事例プール P の要素数は 1500 である．プールにある事例を信頼度に応じてラベル付き事例として採用する数 (一回当たり)n, p は，初期事例の数によって，25 個 (初期事例 10000 の場合)，ならびに，30 個 (初期事例 20000 の場合) とした。また，Co-training の繰返し回数は，初期事例 10000 の場合が 400 回，初期事例 20000 の場合が 200 回である．

4.2 各方法の比較実験および考察

節 3.4で述べた，素性集合間の共起性を考慮する手法を「共起」と表し，素性頻度を均一化するようにランダ

ムに振り分ける手法を「頻度均一」と表すものとし，これらの比較検討を初期事例 10000 の場合と 20000 の場合について行なう。まず，Co-trainig における，各学習器 h_1 ，h_2 の精度の変化について表 1,2,3,4に示す．

表 1: Co-training 前後における各学習器の精度の変化 (手法: 共起，初期事例 10000)

	分類器 h_1		分類器 h_2	
Co-training	前	後	前	後
再現率	45.2	47.3	53.7	53.8
適合率	69.8	66.1	82.4	82.1
F 値	54.9	55.2	65.0	65.0

表 2: Co-training 前後における各学習器の精度の変化 (手法: 頻度均一，初期事例 10000)

	分類器 h_1		分類器 h_2	
Co-training	前	後	前	後
再現率	62.1	62.9	40.4	40.1
適合率	69.3	72.2	79.2	78.5
F 値	65.5	67.2	53.5	53.1

表 3: Co-training 前後における各学習器の精度の変化 (手法: 共起，初期事例 20000)

	分類器 h_1		分類器 h_2	
Co-training	前	後	前	後
再現率	63.4	63.3	50.2	50.7
適合率	83.5	82.8	89.7	89.3
F 値	72	71.7	64.4	64.7

表 4: Co-training 前後における各学習器の精度の変化 (手法: 頻度均一，初期事例 20000)

	分類器 h_1		分類器 h_2	
Co-training	前	後	前	後
再現率	66.4	67.0	42.7	42.8
適合率	85.5	86.0	91.6	91.0
F 値	74.8	75.4	58.2	58.2

また、 Co-training により新たに得られた事例を初期事例に加えたものを学習事例として，素性集合の分割を行なわずにすべての素性をつかって学習した場合の精度を，ない) 精度を計り直してみると次のような結果が得られた。表 5,6に示す．

上記結果をみると，まず，手法「頻度均一」のほうが手法「共起」よりも Co-trainig の効果が現れていることがわかる．初期事例数 10000 の場合には，わずかではあるが，全体の学習精度として再現率，適合率の両者が上昇していることがわかる．初期事例数 20000 の場合には，全体の学習精度として適合率の低下を抑えつつ，再現率を上昇させていることがわかる．採用事例を少なくし，ラベル推定の精度を上昇させれば，さらなる効果が得られるかもしれない．ただし，各学習器の精度のバランスが良くないので，素性集合の選択には，まだ検討の余地がある．

5 まとめと今後の課題

本稿では，SVM と Co-trainig に基づく固有表現抽出器の学習について検討を行ない，本手法で Co-training

表 5: 獲得された新規事例を加えた学習精度 (初期事例 10000)

	元の用訓練事例	共起	頻度均一
再現率	63.3	63.3	63.6
適合率	84.7	84.7	85.9
F 値	72.5	72.5	73

表 6: 獲得された新規事例を加えた学習精度 (初期事例 20000)

	元の訓練事例	共起	頻度均一
再現率	70.6	70.5	71.5
適合率	88.4	88.5	88.3
F 値	78.5	78.5	79.0

が可能であることが確認された．しかし，本実験では，Co-training の効果があまり顕著に見られていない．その原因としては，いずれの場合も，採用事例の精度が十分に高くないということが観察されていることが挙げられる．これは，一度に採用する事例数がまだ多すぎることが原因と考えられる．また，左向き解析を考慮した我々の手法では，学習初期段階とそれ以降では追加すべき事例のラベルの頻度分布が変わってくる．今後，これらの要因を検討し，考察を重ねたい．

参考文献

[1] 山田 寛康、工藤 拓、松本 祐治. Support Vector Machines を用いた日本語固有表現抽出. 情報処理学会自然言語処理研究会 Vo.2001,No.20. NL-142-17, pp.121–128 2001.

[2] 山田 寛康、松本 祐治. Support Vector Machine の多値分類問題の適用法. 情報処理学会自然言語処理研究会 Vo.2001,No.112. NL-146-6, pp.33–38 2001.

[3] 新納 浩幸. 素性間の共起性を検査する Co-training による語義判別規則の学習. 情報処理学会自然言語処理研究会 Vo.2001,No.86. NL-145-5, pp.29–36 2001.

[4] Michael Collins and Yoram Singer. Unsupervised Models For Named Entity Classfication. Joint SIGDAT Conference on Empirial Methods in Natural Language Processing and Very Large Corpora. pp.100–110 1999.

[5] Avrim Blum and Tom Mitchell. Combining Labeled and Unlabeled Data with Co-training. The Proceedings of the 1998 Conference on Computational Learning Theory. 1998.