修　　士　　論　　文

知識ベースを利用した教師あり薬物間相互作用抽出の改善

Improving Supervised Extraction of Drug-Drug Interaction Utilizing Knowledgebases

指導教員　　近山 隆 教授

東京大学工学系研究科
電気系工学専攻

氏 名　**37-126468** 成川 弘樹

提 出 日　　平成 **26** 年 **2** 月 **6** 日

概 要

テキストからの情報抽出は，情報検索や知識ベースの構築において重要なタスクである．エンティティ同士の関係についての言及は文の情報の中で重要なものであることがあるため，関係抽出は情報抽出を行うのに必要な作業である．生物医学分野の文献においては，薬物間，また蛋白質間の相互作用の情報は重要な情報であり，これを抽出する試みは広く行われている．

関係抽出においては機械学習を用いた手法が多くなされている．教師あり学習を利用した研究も多くなされているが，教師データとするために正解ラベルを付すことは人的コストが大きい．そのため，正解ラベルを付されていないテキストを大量に使用し，知識ベースを援用する，distant supervision と呼ばれる手法で関係抽出を行うことがなされている．New York Times と FreeBase を用いた distant supervision はニュース記事からの関係抽出で大きな成果を上げている [9]．この手法は薬物間相互作用抽出においても多く研究されているが，薬物間相互作用に distant supervision を適用した研究では，教師あり学習を用いた研究とくらべてより多くの学習データを用いているにも関わらず，大幅に精度が低いものとなっている [22]．

この研究では，ラベル付きデータに知識ベースを利用することにより自動的にラベル付けしたデータを追加することによって精度を向上する手法を提案する．評価実験として，ラベル付きデータによる学習と，それにラベルなしデータに知識ベースによる付したラベルを追加したデータによる学習を行い比較した．その結果，ラベル付きデータのみで学習する場合と比べ，知識ベースによるラベルを適切な重みで付した場合，Precision-Recall 曲線の Area Under the Curve で 56.6%から 65.9%に改善することが確認された．これにより，知識ベースを利用してラベルなしデータに精度の低いラベルを付した場合でも，それらに適切な重みを付した上でラベル付きデータに追加することが，学習精度向上に役立つことが確認できた．

目 次

図 目 次

アルゴリズム，擬似コード

第1章　はじめに

1.1　背景

自然言語を機械で処理することに関する研究は広く行われている．自然言語で書かれたテキストから情報を抽出することは，自然言語の重要なタスクの一つであり，研究が行われている．テキストから情報を抽出することができれば，情報検索において具体的な内容を指定した情報検索が行えるようになるほか，自然言語で書かれた文書を読み込むことで知識ベースを自動構築することもあげられる．文書中に書かれた情報の中でも，文書中に書かれたエンティティ同士の関係についての言及は，文書から抽出すべき重要なものであり，これを抽出する研究は多く行われている．生物医学文献から，蛋白質や薬物の相互作用を抽出する試みもこの関係抽出の一つであり，研究が行われている．

近年では，自然言語処理では多くのタスクで機械学習を用いた手法が主流となっており，情報抽出タスクにおいても多くのテキストを用意して機械学習を行う手法が多くなされている [2, 13, 22, 25]．機械学習においては，学習データを多く用意できれば精度の向上に繋げられる一方，人間によって正解ラベルを付された学習データを用意することは人的コストが高いため，多く用意することが難しい．一方，ラベルの付されていないデータはラベルのついたデータと比べて多く用意しやすい．そのため，情報抽出タスクにおいてもラベルのついていない文書データと知識ベースを利用した研究が盛んに行われているが，ラベルの付されたデータと比べると学習に使用することが容易ではなく，生物医学分野の相互作用抽出においてはラベルの付されたデータを使用した研究とくらべてはるかに多くのデータを学習に使用しているにも関わらずラベルの付されたデータのみから学習した場合と比べて下回っている [22]．

1.2　本研究の目的

本研究では，情報抽出，中でも生物医学分野のテキストから薬物間の相互作用の精度の向上を目的とする．

薬物間相互作用抽出においては，ラベルのついていないデータを多く集めて学習を行った場合であっても，それより少ないラベルのついたデータを用いた学習と比べて精度が向上しないという問題点があった．一方で，ラベルのついていないデータは多く集められるため，このデータの活用は今後の精度向上の観点からも重要であると考えられる．ラベルのついていないデータを利用するにあたっては知識ベースを用いてラベルを付することによってラベルを付することで学習を行う方法が盛んだが，この手法で付したラベルは精度が低く，人手によって付された正解ラベルと同様に扱うことによって精度を出すことは難しい．

1.3 本研究の貢献

本研究では，ラベルの付されていないデータに知識ベースを用いてラベルを付すことで得られる精度の低いラベルを使用することで，ラベルの付されたデータを利用した場合の精度の高いが量の少ないラベルのみから学習する場合と比べて精度を向上することができるようになることを示した．特に，薬物間相互作用においては，ラベルの付されていないデータは多くあるものの，ラベルの付されていないデータと知識ベースを用いた学習でも精度よく抽出を行うことができなかったが，重みを調整することにより，ラベルの付されたデータのみの場合，あるいは重みを調整せずラベルの付されたデータと同様にラベルの付されていないデータを用いた場合よりも精度が出ることを示した．これは，薬物間相互作用を含む関係抽出タスクにおいて有用であると言える．

1.4 本研究の構成

まず 1 章にて本研究の背景や目的について述べる．続いて，第 2 章で，関連研究として，自然言語処理のための機械学習についての研究，および関係抽出に関する研究を挙げる．続いて第 3 章で提案手法について説明し，その評価実験の結果を第 4 章で述べ，それにもとづいて第 5 章にて本研究の結論と今後の課題について述べる．

第2章　関連研究

2.1　関係抽出

自然言語で書かれたテキストにおいて，言及された人や地名，組織といったエンティティの間の関係を抽出するタスクを，関係抽出（Relation Extraction）と呼ぶ．例えば，

(1)　*Hongo campus is located in Tokyo.*

という文は，*Hongo campus* と *Tokyo* の関係を示しており，その関係は「そこに住んでいる」という関係であることがわかる．これを，計算機上で扱い易い形式で，LocateIn(Hongo campus, Tokyo) のように表現するものである．この抽出タスクにあたっては，まずこの文から “*Hongo campus*”, “*Tokyo*” という二つのエンティティを検出し，その二つのエンティティについてどのような関係が記されているのか，またはどのような関係も記されていないのかを機械学習の結果を利用して調べることによってこの関係を抽出することができる．エンティティが 3 つ以上検出された場合には，その中から 2 つを選び出す組み合わせそれぞれについて関係を調べることになる．

2.1.1　薬物相互作用抽出

薬物相互作用（Drug-Drug Interaction）の抽出は，関係抽出の一分野である．これを行うことにより，医学文書に登場する薬物同士の関係を抽出することで，医学文書の情報を整理したり，必要な関係に関する文書を検索したり，究極的には論文から薬物の関係に関する知識ベースを構築することが考えられる．薬物間の関係は，単純に薬物同士で起こるのみならず，体内に同時に存在することにより，身体の作用などを通してお互いの濃度を変化させたり，役割を阻害または促進したりといったものが含まれる．また，それを行うことを奨めていることもあれば，禁忌とされていることもある．また,「同時に投与しても，特に相互作用が起こらない」とする文もある．もちろん，列挙されているだけで特に関係を示さないものもある．このように，一つの薬物ペアの関係についても多くの場合が考えられる．今回の研究では，関係を示しているかどうかのみを考え，具体的にその関係についてどのような情報が付与されているかについては対象としない．また,「関係を示さない」という情報についても関係を示していないものとして扱う．

例えば，

(2)　***Chloral hydrate*** *and* ***methaqualone*** *interact with* ***anticoagulant*** *agents.*

という文があった場合に，*Chloral hydrate* と *anticoagulant*，*methaqualone* と *anticoagulant* はそれぞれ関係を示しており，関係をもつことが記されている，*Chloral hydrate* と *methaqualone* は併記されているだけで特に関係を示すものではない．この場合，入力として文そのものに加え

- (*Chloral hydrate*, *methaqualone*) を与えた時に「相互作用なし」
- (*Chloral hydrate*, *anticoagualnt*) を与えた時に「相互作用あり」
- (*methaqualone*, *anticoagulant*) を与えた時に「相互作用あり」

と返すような関数を作ることが目標となる．すなわち，(文，文中の薬物 1，文中の薬物 2) を入力とし，その二つの薬物が関係しているかが出力となる．文中に含まれる薬物の数 n に対して ${}_nC_2$ 通りの薬物ペアが考えられる．

2.2 教師あり学習

教師あり学習とは，入力データとそれに対する正解が書かれたデータを利用することで学習を行う手法である．

2.2.1 分類器

自然言語処理における機械学習では，文から特徴ベクトルを抽出し，それを元に分類器で分類を行うことで出力を得るものがあり，提案手法でも使用している．二値分類においては入力された特徴ベクトルを $\boldsymbol{x}$ とした時，$f(\boldsymbol{x})$ が一定以上の時とそうでない時の 2 値で分類を行うことが考えられる．これを教師ありによって行う時，学習データ $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_n, y_n)\}$ を用意できればこの学習データを最も良く分類するような関数 f を作ることにより未知の入力値の分類を行うことになる．

線形分類器においては，ベクトル $\boldsymbol{w}$ を用いてこの f を $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x}$ と表現することができる．学習データを入力した時，最適な $\boldsymbol{w}$ を見つけることが目的である．

2.2.1.1 ロジスティック回帰モデルによる分類器

ロジスティック回帰モデルとは，回帰モデルの一つであり，量的変数から，0 と 1 の二値で表現されるような質的変数を予測する際に用いられるものである．ある線形関数 $f = b_0x_0 + b_1x_1 + ... + b_(r-1)x_(r-1) + b$ を予測を行う時，式 (2.1) のようなモデルを立てる．

$$p(\boldsymbol{x}) = Pr(y = 1|\boldsymbol{x}) = \frac{exp(f(\boldsymbol{x}))}{1 + exp(f(\boldsymbol{x}))} = \frac{1}{1 + exp(-f(\boldsymbol{x}))} \tag{2.1}$$

この時，$\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})} = exp(f(\boldsymbol{x}))$ はこの事象が起こる確率の，起こらない確率に対する比率（オッズ）と表現される．

これを $p(\boldsymbol{x})$ と $f(\boldsymbol{x})$ の関係とみなすと，式 (2.2) となる．これは，シグモイド関数と呼ばれ，図 2.1 のようになる．

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.2}$$

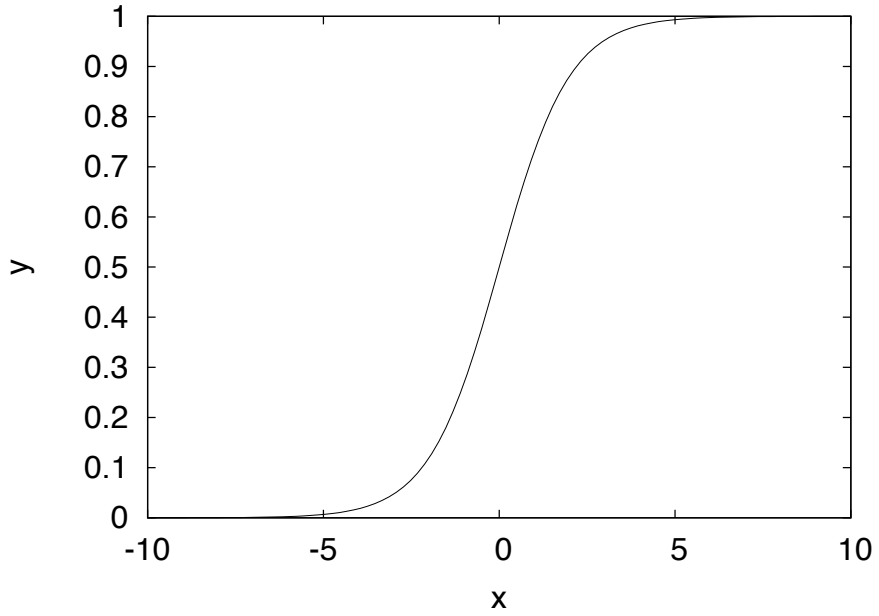


図 2.1. シグモイド関数

この回帰モデルを損失関数として用い，L2 正則化を用いて分類器を学習する場合，最適な分類器を見つける問題は式 (2.3) のように表現される．この時 $\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$ は，正例と負例の間のマージンを最大化するためのものである．これを図 2.2 に示す．

$$\min_{\boldsymbol{w}} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l} \log(1 + exp(-y_i\boldsymbol{w}^T\boldsymbol{x}_i)) \tag{2.3}$$

2.2.1.2 信頼領域法

ロジスティック回帰モデルにおける学習にあたって，実際に式 (2.3) における目的関数を最小化する $\boldsymbol{w}$ を求める手法として，信頼領域法 [12] があり，これは提案手法で用いる LibLINEAR [5] でのロジスティック回帰モデルなどでの学習でも使われている．

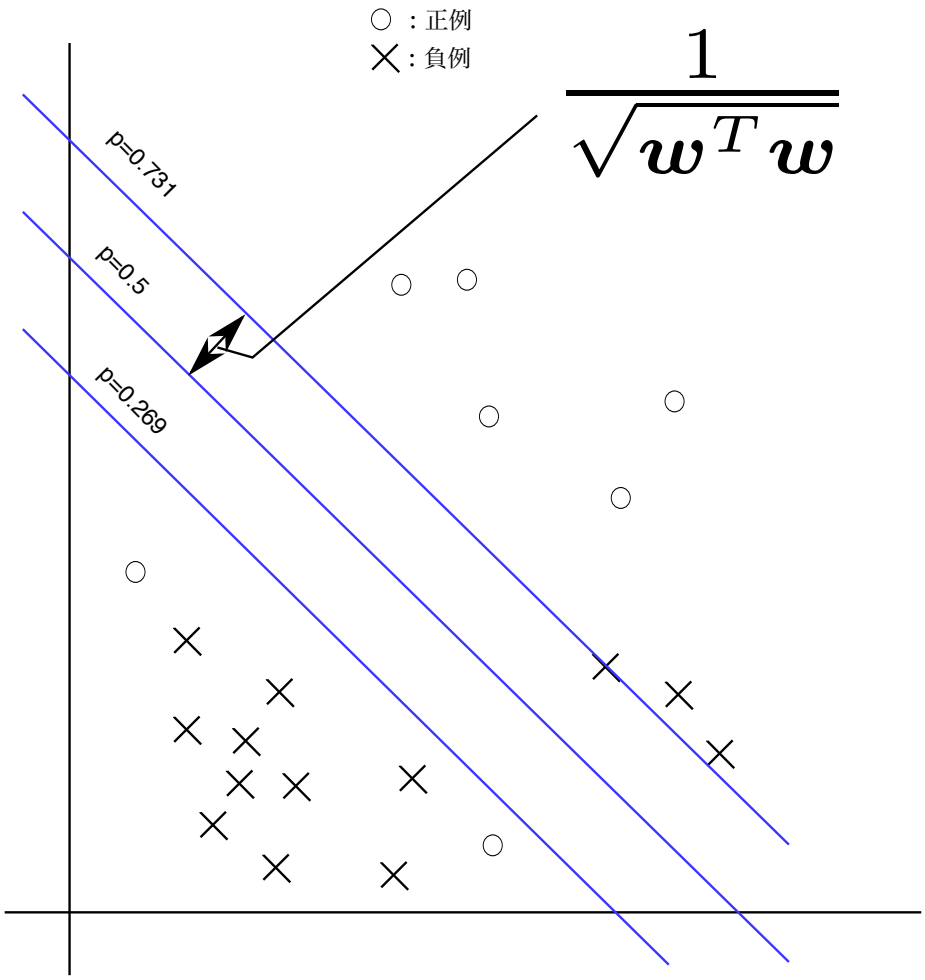


図 2.2. ロジスティック回帰モデルにおける分類

目的関数を $f(\boldsymbol{x})$ とし，これをニュートン法を用いて最小化していくが，その際，ニュートン法のモデルが適用できると考えられる信頼領域を都度計算し，その領域内での最善解を求めてることを繰り返す．アルゴリズムは Algorithm 2.1 のようになる．ただし，$\sigma_1, \sigma_2, \sigma_3, \eta_1, \eta_2$ は $0 < \sigma_1 < \sigma_2 < 1 < \sigma_3$，$0 < \eta_1 < \eta_2 < 1$ な定数である．また，q_k は式 (2.4) で表される．

$$q_k(\boldsymbol{s}) = \nabla f(\boldsymbol{w}^k)^T \boldsymbol{s} + \frac{1}{2} \boldsymbol{s}^T \nabla^2 f(\boldsymbol{w}^k) \boldsymbol{s} \tag{2.4}$$

$$\begin{aligned} &\Delta_{k+1} \in [\sigma_1 \min\{||s^k||, \Delta_k\}, \sigma_2 \Delta_k] \text{ if } \rho_k \leq \eta_1 \\ &\Delta_{k+1} \in [\sigma_1 \Delta_k, \sigma_3 \Delta_k] \text{ if } \rho_k \in (\eta_1, \eta_2) \\ &\Delta_{k+1} \in [\Delta_k, \sigma_3 \Delta_k] \text{ if } \rho_k \geq \eta_2 \end{aligned} \tag{2.5}$$

Algorithm 2.1 信頼領域法のアルゴリズム [5]

$\boldsymbol{w}_0$ を初期値に
for k in 0, 1, ... **do**
  $\nabla f(\boldsymbol{w}) = \boldsymbol{0}$ なら終了
  現在の信頼領域の中での最善の解を得る． $\boldsymbol{s}^k = min_{\boldsymbol{s}} q_k(\boldsymbol{s})$，ただし $||\boldsymbol{s}|| \leq \Delta_k$
  $\rho_k = \dfrac{f(\boldsymbol{w}^k + \boldsymbol{s}^k) - f(\boldsymbol{w}^k)}{q_k(\boldsymbol{s}^k)}$
  $\boldsymbol{w}^{k+1} = \boldsymbol{w}^k + \boldsymbol{s}^k$ if $\rho_k > \eta_0$, else $\boldsymbol{w}^{k+1} = \boldsymbol{w}^k$
  式 (2.5) に従って Δ を更新
end for

2.2.2 分類問題における評価指標

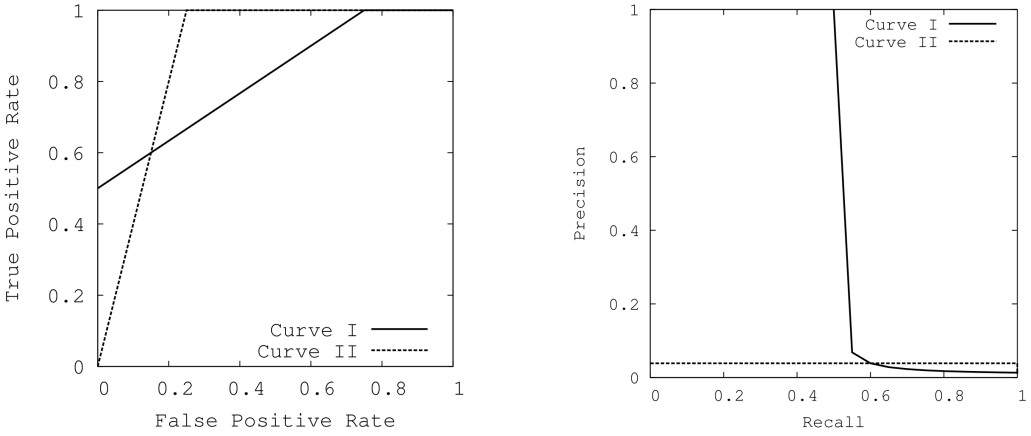
適合率と再現率を評価手法と用いるにあたり，より正例である可能性が高いと判断される例のみを正例とすることにより，再現度は下がるが適合率を上げることができる．そのため，分類器についてこの 2 つの値の関係をとることができる．縦軸，横軸それぞれを適合率，再現率としたグラフを Precision-Recall (P-R) 曲線と呼ぶ．

一般に，適合率を高めようとすれば，より確実なデータのみを抽出することとなり，再現率は低下する．一方で，再現率の高い分類を行おうとする場合は，多くのデータを正例としてラベル付けする必要があり，適合率は低下する傾向にある．この二つはトレードオフの関係にある．現実世界においての問題では，正例を誤って負例とすることの問題と，負例を誤って正例とすることの問題が異なることがあり，一つの点を以て評価することは分類器の性能指標として適切でないことがある [19]．例えば，マルウェアや詐欺，火災など，問題が発生した際にそれを検出するシステムにおいては，負例を誤って正例と分類する，すなわち誤報が発生することよりも，正例を誤って負例と分類する，す

なわち検出に失敗することの方が問題が大きいため，適合率が許容できる範囲にとどまっていれば再現率が高い方がよいシステムと言える．一方，特定の知識を検索するプログラムにおいては，その全てを抽出しなくともよく，抽出されたものの精度，すなわち適合率が高い方が評価が高い場合もある．あるいは，マルウェア検出に近いものであっても，スパムフィルタのようなシステムにおいては誤検出のコストが大きいこともある．また，同じシステムにおいても，使用される場所それぞれにおいてどちらが重視されるかは異なる．どちらに重点を置くかに関わらず，分類器によって精度の高い分類を行うこと，すなわち，正例はより高いスコアを，負例はより低いスコアを得ることは，分類器の性能を評価するにあたって重要である．

また，この曲線全体を見た指標として，曲線の下の面積である Area Under the Curve(AUC) が用いられる．

AUC を用いるにあたり，正例と比べて負例が大幅に多いようなデータセットを用いる場合，ROC を用いた場合では正しく分類された多くの負例に影響されて分類器の性能の差が数値としてあまり現れないことがあり，そのような場合でも P-R 曲線においては分類器の性能差が現れやすい [4]．図 2.3 は [4] に掲載されている二つのグラフであるが，ROC の曲線ではその両者の性能差が大きく現れないのに対し，P-R 曲線では再現率を下げ適合率を優先した際の性能において大きな差があることがわかる．



(a) Comparing AUC-ROC for two algorithms

(b) Comparing AUC-PR for two algorithms

図 2.3. ROC と P-R 曲線の両者での同一の 2 つの分類器の比較 [4]

2.2.3 教師あり学習による関係抽出

教師あり学習を利用して関係抽出を行った研究について挙げる．[7] では，分類器として SVM を利用し，素性として

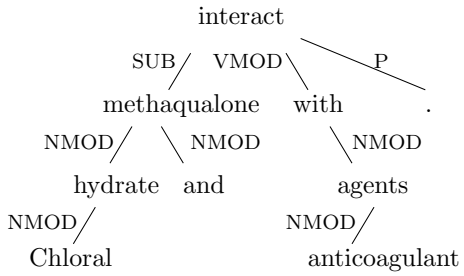
- 単語からの特徴
  - それぞれのエンティティの bag-of-words
  - それぞれのエンティティの head word
  - それぞれのエンティティの head words の組み合わせ
  - 二つのエンティティの間に単語がない場合のみの素性
  - 二つのエンティティの間にある中で（最初・最後）の単語
  - 二つのエンティティの間にある他の単語
  - 先に出現するエンティティの（1 つ・2 つ）前の単語
  - 後に出現するエンティティの（1 つ・2 つ）後の単語
- エンティティ種別
  - それぞれのエンティティ種別（人/団体/設備/場所/地政学上の場所のいずれであるか）の組み合わせ
- 言及のされ方
  - それぞれのエンティティの言及のされ方（名前/名詞句/代名詞のいずれであるか）の組み合わせ
- overlap
  - 二つのエンティティの間にある他のエンティティの数
  - 二つのエンティティの間の単語数
  - 片方のエンティティがもう片方に含まれている場合のフラッグ
- フレーズ
  - 二つのエンティティの間にフレーズがない時のみの素性
  - 二つのエンティティの間のフレーズがひとつならそのフレーズ
  - 二つのエンティティの間のフレーズが複数ある時は（最初・最後）のもの
  - 二つのエンティティの間のフレーズのうちここまでで登場しなかったもの
  - 先に出現するエンティティの（1 つ・2 つ）前のフレーズ
  - 後に出現するエンティティの（1 つ・2 つ）後のフレーズ

- 二つのフレーズの間のフレーズの（そのもの・head word）の並び

- 依存木
  - それぞれのエンティティの種別とその直接の依存先の単語の組み合わせ
  - それぞれの head word とその直接の依存先の単語の組み合わせ
  - 二つのエンティティ種別両方と，それぞれに共通の依存先の（名詞句・形容詞句・動詞句）の存在の有無

を利用することにより，分類を行っている．ただし，この一覧で，括弧で挟まれた中に複数の要素が「・」で区切られている時はそれぞれについての素性があることを示す．

また，一覧に「依存木」とあるのは，依存文法を木で表現したもののことである．図 2.4 は “*Chloral hydrate and methaqualone interact with anticoagulant agents.*” を依存木の形で表現したものである．依存文法では，図 2.4 のように文法構造を表現する．すべての単語は直接または間接に動詞に依存するという考えに基づいて構成されたものである．図 2.4 のような木構造として表現した時，子ノードに該当する単語はその親ノードに依存する単語とされる．そのため，root ノードの単語以外の全ての単語は root ノードの単語に直接または間接的に依存しているとみなされる．



“*Chloral hydrate and methaqualone interact with anticoagulant agents.*”

図 2.4. 依存木の例

生物医学分野の文献からの蛋白質間・薬物間相互作用抽出においても，正解を付したデータがある場合は，それを利用して相互作用抽出を行うことができる [13, 25].

2.3 Distant supervision による学習

教師あり学習では，学習用データに正解ラベルを人の手で付す必要があるため，多くの学習用データを用意することが困難であった．正解ラベルの付された学習データを用いず，正解ラベルの付され

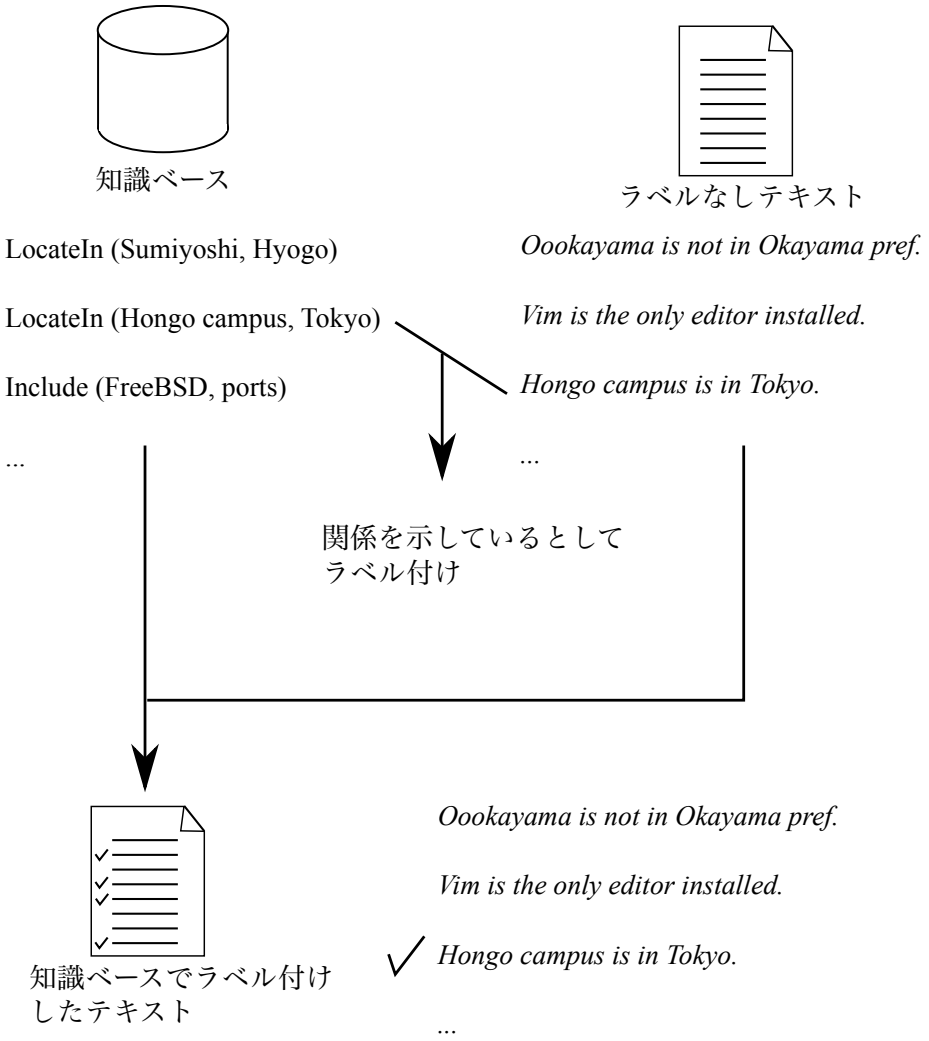


図 2.5. Distant supervision によるラベル付け

ていない学習データを利用することができれば，学習データそのものはより多く用意することができる．そのため，関係についての一般的な知識をデータベース化した知識ベースを利用することによって，正解ラベルの付されていない学習データをを利用した学習を行う，distant supervision による手法が研究されている [9, 14, 15]．2011 年の Hoffmann らの研究 [9] では，複数の関係の種類を定義し，関係について知識ベースで言及されているエンティティペアを含む一つの文はその二つのエンティティの間についての関係が知識ベース上に記されている関係のうちいずれか一つを示しているか，あるいはいずれも示していないとして，また知識ベースによってどんな関係も記されていない関係については関係を示していることはないというモデルをもとに学習を行っている．この条件のもと，複数のエンティティを含む文それぞれについて，定義されたうちのいずれの関係を示しているのか，あるいはいずれの関係も示していないのかについてのラベル付けについて最大確率となるものに更新しながら学習を行うことによって，ラベルの付されていない学習データから関係抽出のパラメータを学習している．

この研究では関係抽出タスクにおいて，

- あるエンティティペアの関係が知識ベースに記されていれば，そのペアを含む文のうち少なくともひとつはその関係を示している
- そうでなければその関係を示している文はない

というモデルを用い，学習の際に実際に関係を示している文の選択を確率モデルを利用することによって行い，New York Times を学習データとして一般ドメインのテキストから学習を行い，文からエンティティ間の関係を抽出できることが示されている [9]．このモデルでは，関係の種類がエンティティペアに対して一つであり，かつそのエンティティペアについての文が他にない場合には図 2.5 のようにラベル付される．ただし，エンティティの抽出については事前に別途行う必要がある．エンティティを精度よく抽出するには，機械学習を利用する手法で抽出する手法がある [6]．一方で，生物医学分野の文献から蛋白質や遺伝子，薬物といったエンティティを抽出する研究では，辞書を元にしたルールベースでの手法が成果を挙げている [8, 11]．

この研究では一般ドメインの自然言語テキストを対象としており，エンティティ間の関係についてあるかないかだけではなくどのような関係があるかについてまでラベル付けするものとし，また一つのエンティティペアについて複数の関係を持つことを想定している．一つの文は含まれる一つのエンティティペアについて最大一つまでの関係を示し，また知識ベースにある関係それぞれは学習データ全体のどこかで言及されている，というモデルを想定し，その制約条件と，それぞれの文がそれぞれの関係を示す確率をモデル化したものを利用することで，確率を最大化するラベル付けを決定し，それを次回以降の学習に利用することとなる．

生物医学分野においても，蛋白質間，また薬物間の相互作用の抽出において，distant supervision を利用した研究はなされている [2, 16, 22, 23]．2012 年の Thomas らの研究 [22] では，蛋白質間，及び薬物間の相互作用の抽出において

- 知識ベースに関係が登録されている組み合わせであれば，その二つのペアに言及した文は全てその二つのエンティティの関係を記している

- そうでなければ二つのエンティティの関係を記していない

というモデルでラベル付を行っている．このラベル付の様子は図 2.5 のようになる．知識ベースとしては，薬物間相互作用においては DrugBank [24]，蛋白質間相互作用においては IntAct [10] などが利用可能である．この手法によりラベル付を行い，それを利用して学習を行うことによって関係抽出を行っている．しかし，Thomas らの研究においても，23,811 エンティティペアからの教師あり学習による教師あり学習による成果の F 値は 62.1%としているのに対し，distant supervision によりラベルを付された 200,000 のエンティティペアを利用して学習を行った場合の F 値はもっとも良かった手法で 9.5%と報告されており，distant supervision による手法では高い精度を出すことが難しいことを示している．

2.4 半教師あり学習

教師あり学習ではラベルの付されたデータを用意するコストが大きく，多くのデータを用いて学習することが困難であった．前節では，ラベルのついていないデータに知識ベースなどを利用して一度ラベルを付し，あるいは学習途中にもラベルを再計算しながら学習を行う手法について述べたが，ラベルの付されたデータがある場合は，ラベル付きデータを利用してまだラベルの付されていない学習データにもラベルを付して学習データとして利用する半教師あり学習と呼ばれる手法がある．この手法においても，ラベルの付されていない多くのデータを利用して学習を行うことができる．

2.4.1 ラベル伝搬アルゴリズム

ラベル伝搬アルゴリズムにより，半教師あり学習を行う手法がある [27]．この手法では，ラベル付きデータとラベルなしデータを利用し，ラベル付きデータから距離の近いラベルなしデータにラベルを伝搬させていくことにより，ラベルなしデータにラベルを付し，多くのデータを利用して学習を行う．基本的な考え方は図 2.6 で表現される．

l 個のラベル付きデータと u 個のラベルなしデータがあり，それぞれの入力は $\boldsymbol{x}_1..\boldsymbol{x}_l$，$\boldsymbol{x}_{l+1}..\boldsymbol{x}_{l+u}$ で定義されるものとする，その時，このアルゴリズムでは，二つの入力のベクトル $\boldsymbol{x}_i$ と $\boldsymbol{x}_j$ が与えられた時，その二つの距離を $d_{ij} = \sqrt{\sum_{d=1}^{D}(x_i^d - x_j^d)^2}$ とし，これが小さいものほど近いラベルを付されるようにすることでラベルの伝搬を行う．$\boldsymbol{x}_j$ からの，それぞれのインスタンス $\boldsymbol{x}_i$ にラベルを伝搬させる重みは $w_{ij} = exp(-\frac{d_{ij}^2}{\sigma^2})$ とする．ただし，σ は重みを調節するパラメータである．この時，伝搬行列 T は式 (2.6) として定義される $(l+u) \times (l+u)$ の行列である．

$$T_{ij} = P(j \to i) = \frac{w_{ij}}{\sum_{k=1}^{l+u} w_{kj}} \tag{2.6}$$

また，分類を行うクラスの数 C に対して $(l+u) \times C$ の行列 Y も用意する．i 行目は $\boldsymbol{x}_i$ に対するそれぞれのクラスについての確率分布を示しており，ラベル付きデータについては付されたラベルのものについて 1，その他について 0 となる．ラベルの付されていないデータについては重要ではない．

この時，アルゴリズムは Algorithm 2.2 のようになる．ただし，計算量を見積もるにあたっては，T の大きさが $(l+u) \times (l+u)$ であることに注意する．

Algorithm 2.2 ラベル伝搬法のアルゴリズム [27]

```
T を計算する
Y を初期化する
while Y が収束していない do
  Y ← YT
  Y を列ごとに正則化
end while
```

このアルゴリズムは，繰り返し数が一定だとしてもインスタンス数 $(l+u)$ に対して計算量が $O((l+u)^2)$ あり，データ量を増加すると急速に計算時間がかかるようになるため，ラベルなしデータを多く得られたとしても，それらのデータ全てを利用して学習を行うことが難しいと考えられる．

2.4.2 Co-training

半教師あり学習の手法として，Blum らによって提案された co-training と呼ばれる手法が挙げられる [1].

Co-training による学習では，2 つの独立していて単独である程度分類が行える 2 つの view を利用するし，図 2.7 のように分類を行う．それぞれで教師あり学習を行い，それぞれの学習結果を元にラベルの付されていないデータにラベルを付すことによって，それぞれの view で片方の view での学習結果を利用して学習を行うことができる．そのため，それぞれの view について，ラベル付きデータを元に高い精度でラベルを付すことができないデータに対してもう片方の view からラベルを付すことができる．この考え方は図 2.8 で表現される．

この二つの view は独立している必要がある． [1] の研究では，ウェブページの分類において，二つの view を

- ページそのものからの view
- そのページのリンク元からの view

としている．また，アルゴリズムは Algorithm 2.3 のようになる．

Blum らによると，この手法でウェブページの分類問題において教師あり学習より良い成果を得られたとしている [1]．また Nigam らの実験でも，記事の分類において教師あり学習と比べて精度が改善されることが確認されている [17].

2.4.3 半教師あり学習による薬物間相互作用抽出

教師なし学習ではデータにラベルが付されいない状態で学習を行うため，学習に使用可能なデータが多くとも一つ一つの精度への寄与は少ない．そのため，少しのラベル付きデータと大量のラベ

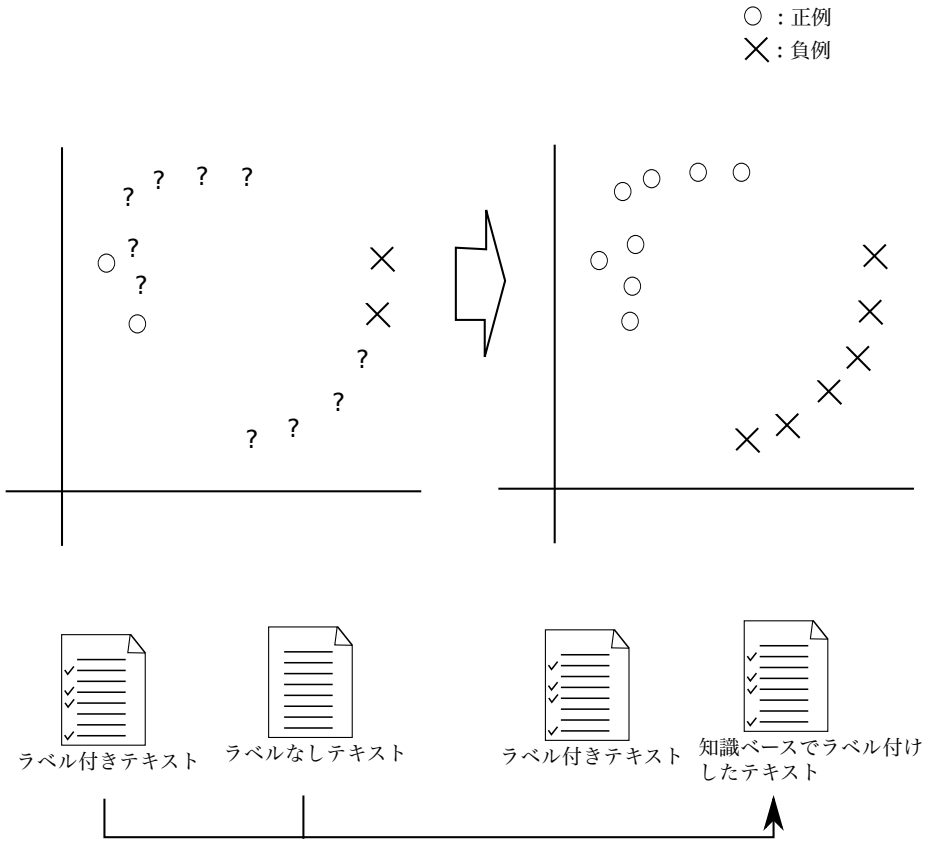


図 2.6. ラベル伝搬法によるラベル付け・学習

Algorithm 2.3 Co-training のアルゴリズム [1]

L をラベル付きデータとする
U をラベルなしデータとする
U から u 個の例をランダムに取り出し U' に移動する
ランダムに u 個の例を U から選択する
for *iteration* in 1..k **do**
　分類器 h_1 を L の view 1 のみで訓練する
　分類器 h_2 を L の view 2 のみで訓練する
　h_1 を利用して U' から p 個の正例と n 個の負例を選択
　選択された $2p + 2n$ 個の例を L に追加する
　$2p + 2n$ 個をランダムに U から選択肢 U' に移動する
end for

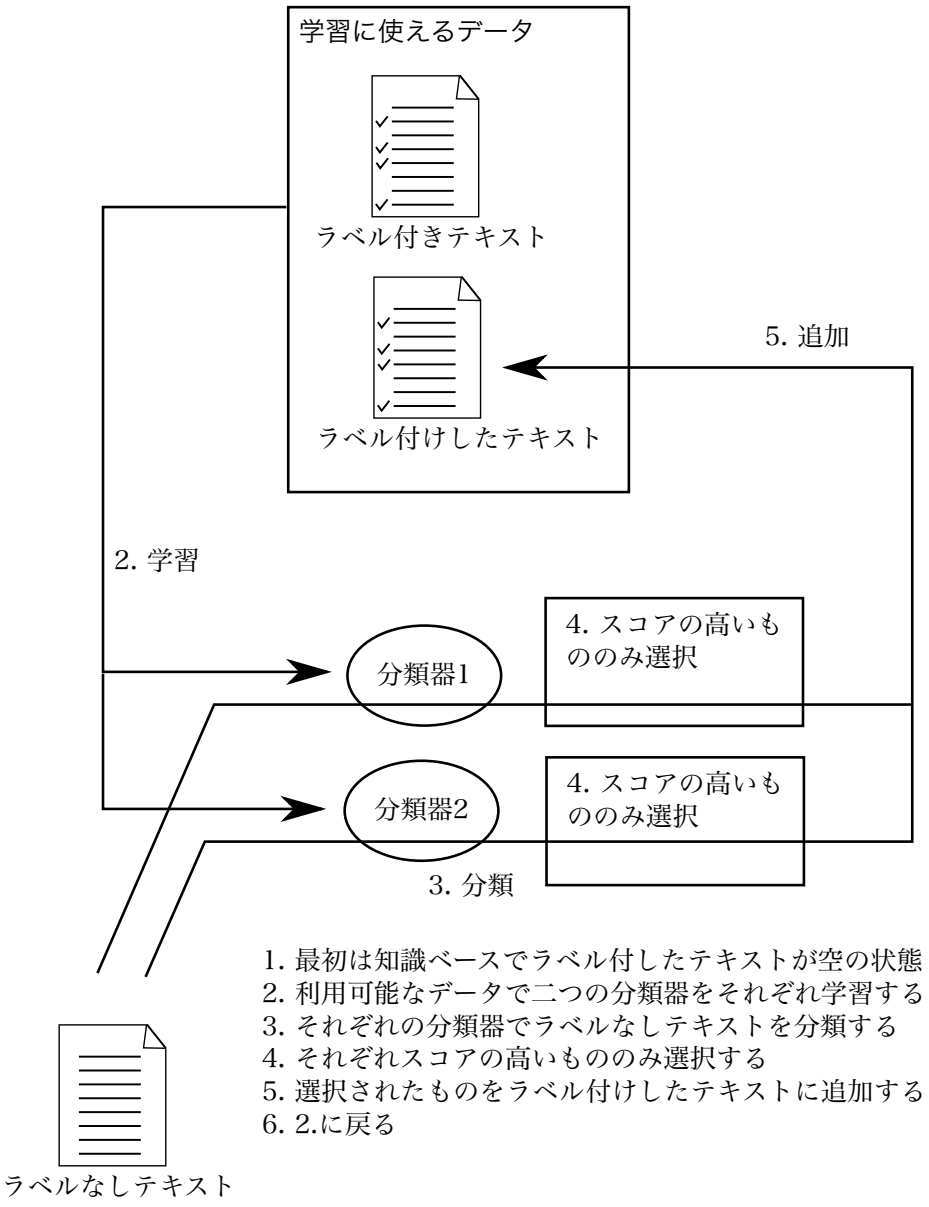


図 2.7. Co-training によるラベル付け・学習

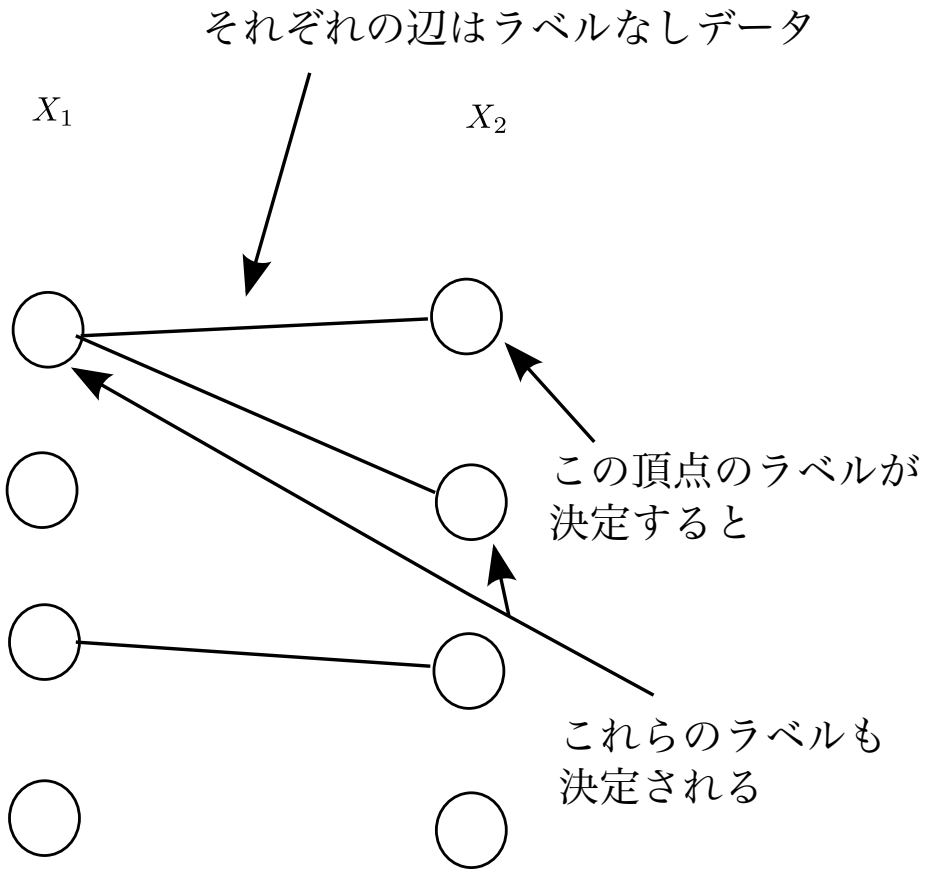


図 2.8. Co-training の考え方 [1]

ルなしデータを用いることにより学習を行うことが考えられる．これを半教師あり学習という．関係抽出タスクにおいて，ラベルありデータとラベルなしデータを利用することにより，学習を行う手法もある．Chen らの実験では，テキストコーパスとしてニュースを用い，一般ドメインからの関係抽出にあたってのラベル伝搬アルゴリズムの有用性が示された [3].

また，蛋白質間の関係抽出タスクにおいて，SVM を用いて分類を行う手法として，教師あり学習を行ったデータを用いてラベルなしデータにラベル付けを行い，一定以上の確率を持って正しいと言えるものをラベル付きデータとして学習を行うことを繰り返す手法によって，教師あり学習を行った場合と比べて高い精度を出すとされている [21]．また，この手法は，k-means 法によってラベルなしデータを分類することによってラベルを付する手法と並ぶ精度を出したとしている [21].

また，Zhang らによる，半教師あり学習による関係分類に関する研究 [26] では，co-training [1] をベースとしながらも，Blum らにより co-training の二つの view の条件として示されている

- それぞれで充分に分類を行える
- ラベルが条件的に独立している

の二つの条件を緩和し，素性として用いられる特徴の次元の中からランダムに分離した複数の view を用いることによって半教師あり学習を行う，BootProject という手法が提案され，実際に精度向上が確認されている．この手法では，それぞれの view について，それぞれの特徴を確率 p（この研究では 0.5）で選択することで，全体の特徴の次元数 F に対して平均して pF の特徴を持つものを複数（この実験では 10）作ることで学習を行っている．また，この手法でラベルなしテキストにラベル付けをするにあたっても，一つの view が高い確度を以てラベル付けすればそのラベルを付した co-training と異なり，複数の view のうち定められた数（この論文では 10 のうち 9 とするのが最も精度向上に寄与したとしている）が一致するラベルを付したものを選択してそのラベルを付している．

BootProject のアルゴリズムは Algorithm 2.4 の通りである．

2.5 パターン抽出による関係抽出

前節までで素性を抽出することにより，機械学習を用いて関係を抽出する手法について述べてきたが，異なるアプローチとして，関係を示すパターンを抽出し，そのパターンを検出することによって関係を抽出する研究がある．関係を示す文に頻出するなるべく一般的なパターンを抽出することにより，適合率が低くなるものの再現率の向上を目指す Pantel らの研究 [18] においては，関係を示すエンティティペアを最初にいくつか用意し，

- パターン生成: 関係を示すエンティティペアを含む文を学習データから抽出し，一般的なパターンを生成する
- パターンランキング・選択: 一定以上の信頼度のパターンを抽出する
- インスタンス拡張: 新たに生成されたパターンを含む文を抽出し，関係を示すエンティティペアの一覧を増やす

Algorithm 2.4 BootProject のアルゴリズム [26]

L をラベル付きデータとする
U をラベルなしデータとする
バッチの大きさを S とする
view の数を P とする
それぞれの特徴が選択される確率を p とする
repeat
  for i in 1..P **do**
    それぞれの特徴を確率 p で選択することで，特徴空間 F_i を作る
    F_i の空間上での L と U の写像である L_i と U_i を作る
    分類器 C_i を L_i で学習する
    C_i で U_i 上の各インスタンスを分類する
  end for
  U から，P 個の分類器のうち多くの（事前に定めた最小数は下回らない）分類器が同じラベルを付したもの最大 S 個にラベルを付す
  それらを L に追加する
until ラベルなしデータが残されていないか，新たにラベル付けを行うに至る合意を得られたデータがない

の 3 つの工程を繰り返すことでパターンを生成している．また，この手法により関係を抽出できる．

第3章　知識ベースによる教師あり学習の改善

関係抽出において，教師あり学習を用いる手法は成果を挙げているが，人の手でラベルを付すことはコストが大きく，多くの学習データを利用して学習を行うことが難しいという問題があった．一方，知識ベースをもとにラベルを付す distant supervision により学習を行ったものは，より多くの学習データを用いているにも関わらず，ラベル付けの精度が低いため，薬物間相互作用抽出においては教師あり学習と比較して劣る結果となっている．つまり，

- 学習データの数（量）
- 学習データに付されたラベルの精度（質）

がトレードオフとなり，特に薬物間相互作用における知識ベースからのラベル付では後者の影響が大きく，精度が出ていないという問題があった．

半教師あり学習では，質の高い少量のデータを元にしつつ，新たに数の確保しやすいラベルなしデータにラベルを付すことで両者のメリットを利用しているが，これは教師あり学習のみでも充分な精度でラベル付できることが前提とされており，また，手法によってはラベルを付すことの計算量の多さによりデータを増やすことが難しいという問題もあった．

ここまでで言及したリソースとして

- ラベル付きテキスト（精度が高いが量が確保できない）
- ラベルなしテキスト（ラベル付けをする必要があり，そのため精度の確保も難しいが，量は確保しやすい）
- 知識ベース（ラベルなしテキストのラベル付けにおいて，有意な情報を利用できる）

があった．これらを全て利用することで，教師あり学習の精度を向上する手法について提案する．

3.1　機械学習を利用した関係抽出

本研究では，関係抽出を，文，及びその文中で言及されているエンティティのうち二つを選択したものを入力とし，その二つのエンティティについての関係にその文が言及しているかどうかを出力とする関数として考える．ただし，言及そのものがなされていない場合と，関係がないと言及されている場合は，いずれも区別なく関係について言及していないものとして扱う．エンティティが二つ以上登場するそれぞれの文について，ペアの組み合わせの数だけ入力が発生することとなる．

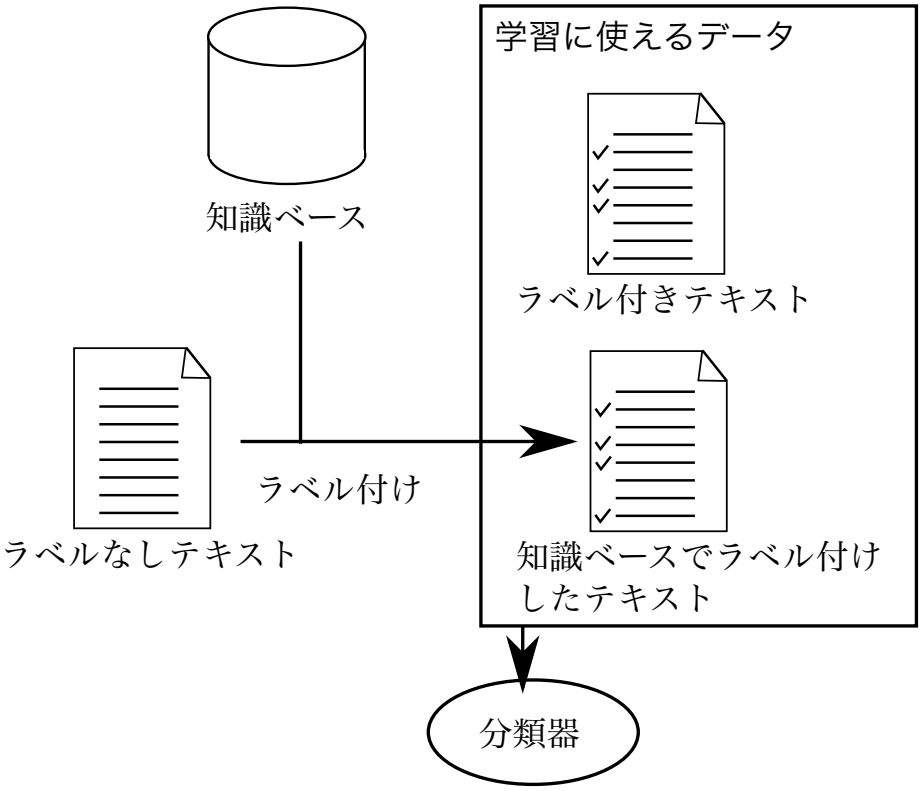

学習に使えるデータ
知識ベース
ラベル付きテキスト
ラベル付け
ラベルなしテキスト
知識ベースでラベル付け
したテキスト
分類器

図 3.1. 学習の流れ

入力と出力が定義され，また，何らかの手段で入力に正解ラベルが付されれば，分類器を利用することによって学習を行うことができる．本研究では，LibLINEAR [5] を用い，ロジスティック回帰モデルによる分類を行った．

3.2 知識ベース

本研究では，薬物そのもの，また薬物間相互作用に関する情報が記されている DrugBank [24] という知識ベースを利用する．この知識ベースでは，薬物そのものに関する情報の他に，それぞれの薬物について相互作用を起こすことが報告されている薬物が登録されている．例えば，表 3.1 はエチルアルコールに関するエントリから一部項目を抜粋したものである．この表に示される通り，今回使用した知識ベースでは，

- その薬物の名称，別名
- その薬物そのものに関する情報
  - 例では説明のみを引用しているが，実際にはより多くの情報が利用可能である．ただし，本研究では使用していない．
- 薬物同士の相互作用として報告されているもの

などが利用可能である．

3.2.1 ラベル付けにおける知識ベースの利用

知識ベースを利用することにより，本研究では，図 2.5 のように，

- 知識ベースに関係が登録されている組み合わせであれば，その二つのペアに言及した文は全てその二つのエンティティの関係を記している
- そうでなければ二つのエンティティの関係を記していない

というモデルに基づき，ラベルなしテキストにラベルを付した．例えば，

(3) ***Ethanol*** *increases the adverse effects of* ***Tizanidine****.*

という文の *Ethanol* と *Tizanidine* の関係についてラベルを付す時，*Tizanidine* は DrugBank に ID: DB00697 で登録されており，表 3.1 に示す通り *Ethanol* との関係が DrugBank に登録されているため，これを正例とラベル付けすることができる．

あくまで知識ベースでのエントリを利用してラベルを付すため，個々の文について正しいラベルを付すことは保証されていない．例えば，開発データ中には次のような文がある．

表 3.1. DrugBank [24] のエチルアルコールの項目（一部項目抜粋，整形）

drugbank-id	DB00898
name	Ethanol
description	A clear, colorless liquid rapidly absorbed from the gastrointestinal tract and distributed throughout the body. It has bactericidal activity and is used often as a topical disinfectant. It is widely used as a solvent and preservative in pharmaceutical preparations as well as serving as the primary ingredient in alcoholic beverages. [PubChem]
synonyms	Absolute Alcohol, Absolute Ethanol, Alcohol, Alcohol Anhydrous, Alcohol, Dehydrated, Alcohol, Diluted, Alcool Ethylique, Alcool Etilico, Alkohol, Alkoholu Etylowego, Aminoethanol, Beta-Aminoethanol, Beta-Aminoethyl Alcohol, Beta-Ethanolamine, Beta-Hydroxyethylamine, Caswell No. 426, Dehydrated Ethanol, Denatured Alcohol, Denatured Ethanol, Etanolo, Ethanol 200 Proof, Ethanol Anhydrous, Ethanol Extra Pure, Ethyl Alcohol, Ethyl Alcohol Anhydrous, Ethyl Alcohol, Anhydrous, Ethyl Alcohol, Denatured, Ethyl Hydrate, Ethyl Hydroxide, Ethylol, Ethylolamine, HSDB 531, Methylcarbinol, USAF EK-1597
drug-interactions	DB01048 name: Abacavir description: Abacavir is partly metabolized through the alcohol dehydrogenase enzyme system. Alcohol increases the area under the curve (about 41%) of Abacavir. Interaction does not appear to be clinically significant. DB00697 name: Tizanidine description: Ethanol increases the adverse effects of Tizanidine. The CNS depressant effects of these agents are additive. DB00427 name: Triprolidine description: Triprolidine may enhance the CNS depressant effects of Ethanol.

(4) *The benzodiazepines, including **alprazolam**, produce additive CNS depressant effects when co-administered with other psychotropic medications, anticonvulsants, antihistaminics, **ethanol**, and other drugs which themselves produce CNS depression.*

この文では，*alprazolam* と *ethanol* の間で相互作用があったことを示されている．*Alprazolam* は，DrugBank の ID: DB00404 として登録されているが，表 3.1 の通り，これと *ethanol* との相互作用は DrugBank に登録されていないため，このラベル付け手法でこの文にラベルを付した場合は相互作用を示していないとしてラベル付けされる．

ただし，文中からのエンティティ抽出については 4.1 節にて述べる．

3.2.2 知識ベースより付与したラベルの精度

この手法で学習用のデータにラベルを付するにあたり，知識ベースを利用して付与したラベルの精度について調査した．開発用データセットにおいて，この方法でラベルを付し，正例と付されたもの，負例と付されたものそれぞれについて，それがどの程度実際のデータと合致するか調べた．開発用データセットのうち，双方のエンティティについて知識ベース上でエンティティを見つけられた 4,092 のエンティティペアについて調査した結果，表 3.2 のようになった．また，これを分類器とみなした時，その精度は表 3.3 となる．

表 3.2. 知識ベースによるラベル付の結果

	実際の関係を示している	実際の関係を示していない	合計
知識ベースに関係がある	153 (34.4%)	292 (65.6%)	445
知識ベースに関係がない	526 (14.4%)	3,121 (85.6%)	3,647
合計	679 (16.6%)	3,413 (83.4%)	4,092

表 3.3. 知識ベースによるラベル付の精度

適合率	34.4%
再現率	22.5%
F 値	27.2%

3.3 ラベル付きデータと知識ベースによるラベルの併用

表 3.2，表 3.3 に示されたように，知識ベースを利用して付したラベルは精度が低い．そのため，distant supervision のみによって学習を行うことによりラベルなしテキストにラベルを付して分類器を学習している研究では，人の手によって付された少数のラベルを用いた研究とくらべても精度が

低い結果が出ている [22]．一方，知識ベースでのデータの有無を参照しているのみのため，確度の高いラベルのみに絞り込むことは困難である．

人の手によって付されたラベルは精度が高く，精度の高い学習を行うことができるが，作成コストが高く，数を多く用意できないため，それほど頻繁に登場するわけではないパターンに対する評価の精度が低くなる，あるいは全くなされなくなる，とする問題点が挙げられる．例えば，”reuptake” の単語については，開発データにおいて 2 エンティティペアしか登場せず，この 2 つは同じ文であり相互作用を示さないものであるが，実際には reuptake は相互作用を示す可能性が高いと考えられる．このように，登場頻度の低いものが多くあると考えられる場合，これらのパターンが学習において正しく作用しない可能性がある．そのため，より精度良い学習を行うためには，データ量を増やす必要があると考えられる．

この問題点を解決するため，学習においてラベル付きデータからの学習を基本的な考え方とするが，ラベルなしデータも知識ベースを利用してラベル付けを行って精度向上に役立てることを考える．両方を分類器の学習に用いるため，学習の流れは図 3.1 に示したように，まずはラベルなしデータに知識ベースを利用してラベルを付し，それをラベル付きデータに加えて分類器の学習に用いる．

3.4　ロジスティック回帰モデルにおける重みの変更

知識ベースを元に新たに付されたデータは精度が低く，人の手で付されたラベルと同等に扱うことは適切でないと考えられる．そのため，本研究では，人の手で付されたラベルと自動的に付されたラベルの重みを変更することにより，精度の向上を図る．

分類器としては，LibLINEAR [5] の L2 正則化ロジスティック回帰モデルをベースとし，知識ベースによるラベルの重みを人手で付されたラベルより下げるため，修正を加えた．式 (3.1) の第 3 項を追加することで，知識ベースによるラベルの重みを人の手で付されたラベルの a 倍とした．

ここで，l は人の手でラベル付けされたエンティティペアの数，u は知識ベースによってラベル付けされたエンティティペアの数である．

$$\min_{\boldsymbol{w}} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\log(1+exp(-y_i\boldsymbol{w}^T\boldsymbol{x}_i)) + aC\sum_{i=l+1}^{l+u}\log(1+exp(-y_i\boldsymbol{w}^T\boldsymbol{x}_i)) \tag{3.1}$$

これにより，ラベルは

- それぞれの重みが 1 である，l 個の，人の手で付されたラベル
- それぞれの重みが a である，u 個の，知識ベースにより付されたラベル

の合計 $(l+u)$ 個で構成されることとなる．

このモデルにおいては，$\boldsymbol{w}$ によって誤って分類されたインスタンスについての損失を，知識ベースで付されたデータであった場合に a 倍としているため，同程度の損失関数の値であるデータであれば，図 3.2 のように知識ベースにより付されたラベルが $\frac{1}{a}$ 個誤分類された時に，手で付されたラベル 1 個が誤分類されたのと同じように扱われる．

そのため，知識ベースによるラベル付けの精度が低かった場合であっても，学習の過程で得られた分類器によって誤分類されることによるペナルティよりも正例と負例の間のマージンの大きさや人の手で付されたラベルの誤分類の影響の方が大きく評価されることになる．

これらにより，提案手法において，本章冒頭で述べた問題点である量と質のトレードオフにおいての両者の利点を利用し，

- 精度の低いラベルにより精度の高いラベルの学習効果が打ち消されることを防ぐこと
  - 精度の高いラベルによる分類，またその分類が誤りとすることによる損失は精度の低いラベルより重視することによる
- 精度の低いラベルを複数組み合わせることで新たな情報を取得し，より精度の高い学習を行うこと
- 知識ベースによって，ラベル付きテキストのみからでは得られない情報を利用した知識の獲得を行うこと

を実現することを期待するものである．

また，本手法において，ラベルなしデータにラベルを付す作業の計算量はラベルなしデータの数 u に対して $O(u)$ であり，学習にあたっては分類器に依存するが，今回使用したロジスティック回帰モデルのみならず，代用としてパーセプトロンなどを利用した場合においても，収束までの繰り返し回数 k と組み合わせて $O(k(l+u))$ であるため，学習データの数を増加させやすいと考えられる．

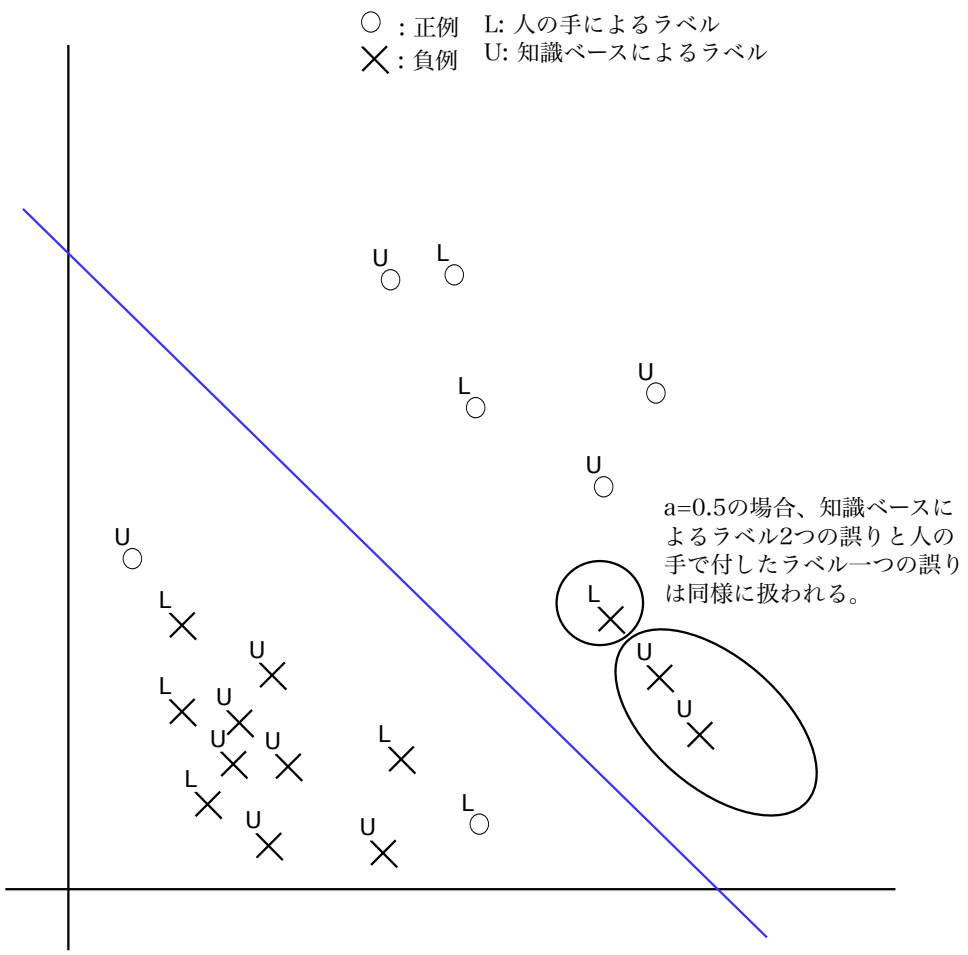

○ : 正例　L: 人の手によるラベル
╳ : 負例　U: 知識ベースによるラベル
a=0.5の場合、知識ベースによるラベル2つの誤りと人の手で付したラベル一つの誤りは同様に扱われる。

図 3.2. 重みの変更

第4章　評価実験

提案手法の有効性について評価するための実験を行った．

4.1　評価設定

4.1.1　データ

今回の実験では，学習データとして以下のデータを用いた．

- ラベル付きテキスト：SemEval 2013 Task 9 用の学習データとして公開されているデータの一部 (11,037 エンティティペア)．ソースが DrugBank と Medline のものがあったため，それぞれの前半部分をとった．なお，これは開発用データセットとして利用した．
  - XML で表現され，個々のエントリは図 4.1 のようになっている．
  - 文中に登場する薬物についてラベル付けがなされている
  - ラベル付けがなされた薬物同士の相互作用が文に示されている場合，薬物を起こす薬物のペアの情報が記されている
- ラベルなしテキスト：PubMed にて提供されている Medline のデータ（3,343,226 エンティティペア）からランダムに 300,000 エンティティペアを抽出
  - エンティティの抽出手法は節にて述べる
- 知識ベース：DrugBank [24] のデータ（6,711 の薬物についてのデータ）
  - その薬物が他の薬物などと相互作用を起こす組み合わせが記されている
  - 「相互作用を起こさない」旨のデータは含まない

また，テキストに関しては GDep [20] を利用して，2.2.3 節で説明した，依存文法の形にパースされたものを利用した．各枝にはどのような関係であるかが示されている．また，原型や品詞などの情報も GDep では利用可能である．本実験では，木の構造と，原型，品詞を利用している．

テストデータとしては，SemEvam 2013 Task 9 用の学習データのうち学習データとして使用しなかった部分（1,970 エンティティペア）を利用した．なお，エンティティ抽出は本研究の目的ではないため，元のデータに存在する存在するエンティティの位置に関する情報はそのまま使用し，それらのエンティティの間の関係抽出の精度についての評価を行った．

```
<sentence id="DrugDDI.d463.s1" text="The steady state plasma concentrations of imipramine a
nd desipramine have been reported to be increased an average of 31% and 20%, respectively,
by the concomitant administration of alprazolam tablets in doses up to 4 mg/day.">
    <entity id="DrugDDI.d463.s1.e0" charOffset="42-51" type="drug" text="imipramine"/>
    <entity id="DrugDDI.d463.s1.e1" charOffset="57-67" type="drug" text="desipramine"/>
    <entity id="DrugDDI.d463.s1.e2" charOffset="182-191" type="drug" text="alprazolam"/>
    <ddi id="DrugDDI.d463.s1.d0" e1="DrugDDI.d463.s1.e0"
        e2="DrugDDI.d463.s1.e2" type="mechanism"/>
    <ddi id="DrugDDI.d463.s1.d1" e1="DrugDDI.d463.s1.e1"
        e2="DrugDDI.d463.s1.e2" type="mechanism"/>
</sentence>
```

図 4.1. ラベル付きデータの例

本研究では，グラフの曲線の下に来る部分の面積である AUC を P-R 曲線に対して計算して分類器の評価尺度とした．

4.1.2 ラベルなしデータからのエンティティ抽出

3.2.1 節で述べた知識ベースの参照においてのエンティティの参照については，知識ベースに登録されたデータを元に抽出を行った．

- ラベルなしテキストから知識ベースの薬物を参照するにあたっては，個々のエントリの name, 及び synonym として登録されているものから一致するものを検索した（表 3.1 参照）
- 名称については正規化を行った
  - 大文字小文字の区別をなくす
  - -と=を同一とみなす
  - 空白を無視する
- 特に誤検出の多かった 23 の名称についてのみ除外対象とした．

4.1.3 素性

本実験では，以下の素性を用いた．いずれも二値のものである．

- 依存木上での 2 つのエンティティの間のパス上での 1-3 gram
- 2 つのエンティティの間にある単語の並びについての 1-3 gram
- 2 つのエンティティの（前・後）にある単語の並びについての 1-3 gram

- 2 つのエンティティの間にある単語の数を離散化したもの
  - 0, 1, 2-3, 4-7, ... というように，二倍になるごとに別の素性のフラグが立つようにした．

ただし，単語の並びについては，単語の原型の並びと品詞の並びをそれぞれ利用した．また，別のエンティティが存在する部分に関しては，それを認識できた場合は全薬物エンティティ共通の文字列に置き換えている．

図 2.4 に示される依存木において，*hydrate* と *agents* の間にあるパスは metaqualone-interact-with となる．ただし，この二つについては方向の関係が存在しないため，逆向きのものに対しても同様に扱った．そのため，この二つであれば metaqualone-interact-with に加えて with-interact-metaqualone も特徴として加えられることになる．

4.1.4 比較対象

提案手法の他に，以下の対象を比較対象として実験結果を記す．

- ラベル付きデータのみを用いた教師あり学習
  - baseline とする手法
  - 本提案手法で追加した，知識ベースでラベルなしデータに対して付したラベルの学習への有用性を評価するため．
- ラベルなしデータと知識ベースのみを用いた distant supervision
  - 関連研究として重要であり，本研究でも重要視されている知識ベースで付されたラベルからの学習の単体での学習効果を確認するため．
- ラベル付きデータとラベルなしデータを用いた，知識ベースからの情報を片方の view とした co-training
  - 同様の手法として考えられる手法についての評価を行うため．

ただし，co-training を行うにあたって，二つの view はそれぞれ

- view 1
  - 文そのものから取得した特徴のみからなる view
- view 2
  - 2 つのエンティティの組み合わせと
  - その 2 つに関する関係の知識ベース上の有無を特徴とする view

の 2 つとした．すなわち，図 2.7 における分類器の片方の入力がエンティティの組み合わせと知識ベースからの入力のみとなる形である．

4.2 評価

評価実験の結果，提案手法，distant supervision，ラベル付きデータのみでの P-R 曲線は図 4.2 のようになった．ラベル付きデータのみで実験した場合と比較して，精度が向上していることが確認できる．また，それぞれの手法での P-R 曲線上での AUC は表 4.1 のようになった．

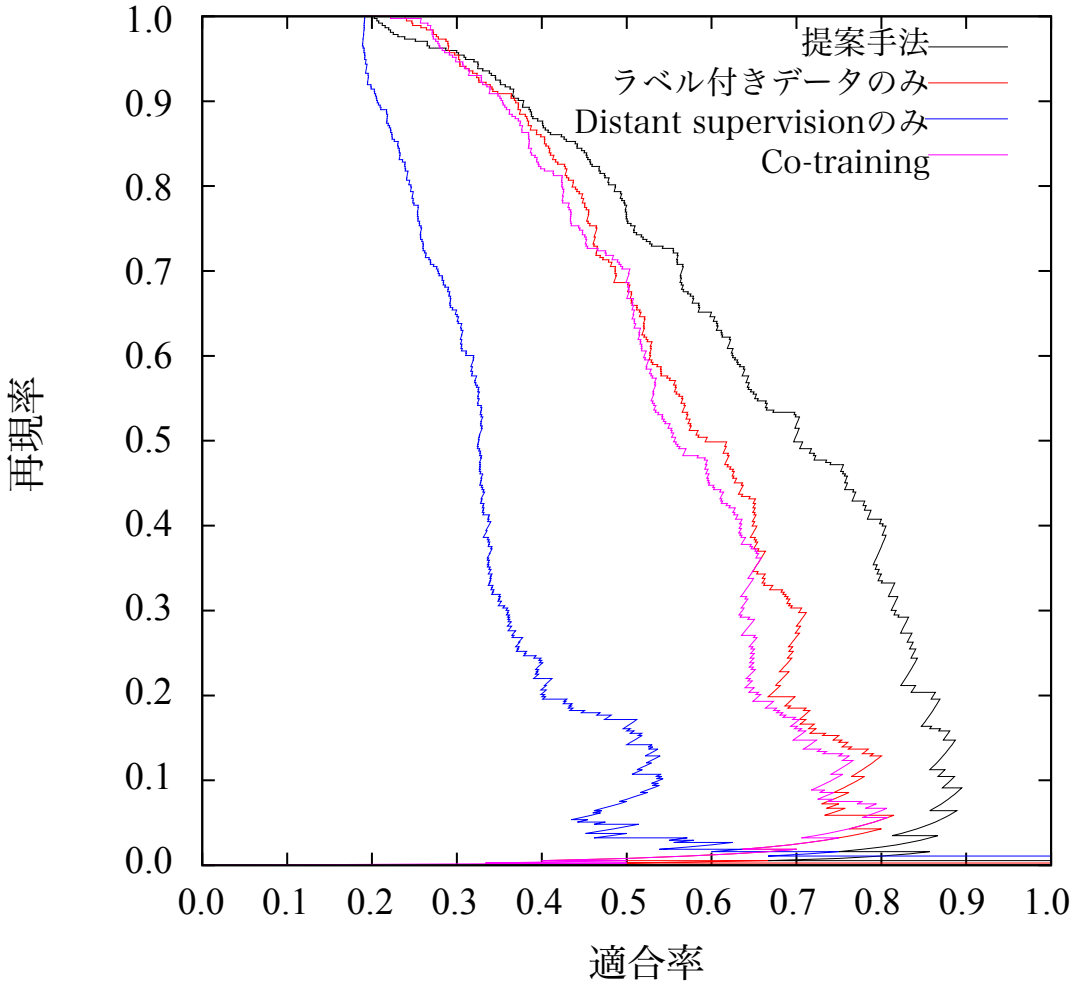


図 4.2. Precision-Recall 曲線

表 4.1. それぞれの手法での AUC

手法	AUC
ラベル付きデータのみ	56.6%
Co-training	54.6%
Distant supervision	33.8%
提案手法	65.9%

4.2.1 提案手法とラベル付きデータのみからの教師あり学習の比較

また，開発用データセットにおいて，交差検定を利用し，知識ベースによるラベルの重みを変化させながら実験を行いそれぞれについての P-R 曲線での AUC を計算したところ，表 4.3 のようになった．ただし，C の値はそれぞれについて調整した．知識ベースにより付されたラベルの重みを手で付されたラベルの 0.01 倍の重みとした時に最大となっている．この条件において，学習に使用できるのは 7,358 のラベル付きデータと，300,000 のラベルなしデータからその時のテストデータとの重複を除いた約 28 万のラベルなしデータである．これの重みは 0.01 倍であるため，手で付されたラベルの重みを 1 とすると，ラベルなし 7,358 に対して知識ベースによって付されたラベルで約 2,800 の重みを有する状態ということができる．つまり，重みで見れば全体の 27%を知識ベースから付したラベルが占めていることになる．

重み（式 (3.1) の a ）の調整を行いながら 300,000 より少ないラベルなしデータを使用して実験を行ってラベル付けを行った場合の精度は表 4.3 のようになり，ラベルを増加させることで精度が向上していることがわかる．

表 4.2. 式 (3.1) の a を増加させた時の P-R 曲線での AUC の変化とその時のパラメータ

式 (3.1) の a	AUC	C
ラベル付きデータのみ	66.7%	20
0.001	76.9%	320
0.002	77.5%	320
0.005	77.4%	80
0.01	77.9%	40
0.02	77.7%	20
0.05	77.3%	10
0.1	76.5%	10
0.2	75.2%	10
0.5	72.5%	2
1	69.9%	2

数字は開発データのもの

また，開発データセット上での交差検定において P-R 曲線を描いた時に，適合率と再現率が一致

する点で，それぞれの分類器が示したラベルごとに数を調べたところ，表 4.4，表 4.5 のようになった．正例，負例ともに，ラベルなしテキストに知識ベースを利用してラベルを付したものにより，より多くのデータを正しく分類することができるようになったことがわかる．

開発データセットにおいて，提案手法とラベル付きデータのみからの教師あり学習の結果を見比べたところ，関係を示す文のうち，ラベル付きデータのみによる学習では抽出できず，ラベルなしデータと知識ベースを利用することにより抽出することができるようになった文としては，二つのエンティティの間の距離が近いもの，また，一文が長いものが目立った．例えば，

(5) *The concurrent use of two or more drugs with anticholinergic activity–such as an antipsychotic drug (eg,* ***chlorpromazine****), an* ***antiparkinsonian*** *drug (eg, trihexyphenidyl), and/or a tricyclic antidepressant (eg, amitriptyline)–commonly results in excessive anticholinergic effects, including dry mouth and associated dental complications, blurred vision, and, in patients exposed to high temperature and humidity, hyperpyrexia.*

という文において，*chlorpromazine* と *antiparkinsonian* の関係の有無の判断を行うにあたって，二つのエンティティの間にある単語はほとんどなく，人間が判断するにあたっても”The concurrent use of...”を見てこの二つを同時使用していることを判断し，更に”*commonly results in excessive anticholinergic effects,...*” を見てその二つの同時使用が相互作用に至ることを理解せねばならない．そのため，多くのテキストを利用できる提案手法においては，少ないデータセットでの学習で抽出することが難しいと思われる素性も多くのデータから取得して利用できると考えられる．また，

(6) *Agents that are CYP3A4 inhibitors that have been found, or are expected, to increase plasma levels of* ***EQUETROTM*** *are the following: Acetazolamide, azole antifungals, cimetidine, clarithromycin(1), dalfopristin, danazol, delavirdine, diltiazem,* ***erythromycin****(1), fluoxetine, fluvoxamine, grapefruit juice, isoniazid, itraconazole, ketoconazole, loratadine, nefazodone, niacinamide, nicotinamide, protease inhibitors, propoxyphene, quinine, quinupristin, troleandomycin, valproate(1), verapamil, zileuton.*

における *EQUETROTM* と *erythromycin* の関係についても，間のテキストではエンティティが多く列挙されており，関係について記されているのはその前にあるテキストであるため，間にあるテキストと比べて素性が取りにくいと考えられる．これも同様に，精度が低くともラベルの付されていないデータを多く用いることにより抽出が行えたものと考えられる．

表 4.3. データ量を増加させた時の P-R 曲線での AUC の変化とその時のパラメータ

ラベルなしデータ	AUC	C	式 (3.1) の a
0	56.6%	20	-
30,000	64.6%	320	0.1
100,000	65.5%	40	0.05
300,000	65.9%	40	0.01

表 4.4. 正例のデータについて二つの分類器が示した出力

		ラベル付きデータのみ	
		相互作用あり	相互作用なし
提案手法	相互作用あり	970	173
	相互作用なし	59	352

表 4.5. 負例のデータについて二つの分類器が示した出力

		ラベル付きデータのみ	
		相互作用あり	相互作用なし
提案手法	相互作用あり	243	170
	相互作用なし	284	8856

また逆に，関係を示していない文のうち，ラベルなしテキストを併用した手法では正しい結果を出したが，ラベル付きテキストのみで学習を行った場合には誤って関係があるとされたデータについて見てみると，

(7) *Although there was no effect of Aprepitant on the plasma AUC of* ***R****(+) or S(-) warfarin determined on Day 3, there was a 34% decrease in S(-)warfarin (a CYP2C9 substrate) trough concentration accompanied by a 14% decrease in the prothrombin time (reported as International Normalized Ratio or INR) 5 days after completion of dosing with* ***Aprepitant****.*

での *R* と *Aprepitant* の関係についても，二つのエンティティの間だけでは決定しづらく，二つのエンティティの外のデータに注目する必要があり，更に，二つのエンティティの間のテキストが長いため，こちらの単語の重要度と外の単語の重要度を正しく判断する必要がある．

一方で，関係を示している文のうち，ラベル付きテキストのみで学習する手法では抽出できていたにも関わらず，ラベルなしテキストを使用して学習を行うと抽出できなくなった関係もある．例えば

(8) *Other 5-HT1B/1D Agonists Concomitant use of other* ***5-HT1B/1D*** *agonists within 24 hours of treatment with* ***AXERT*** *is contraindicated.*

という文においての，太字とした二つのエンティティ間の関係について抽出するにあたっては，間のテキストのみ見ると両方を服用した旨が記されており，素性の重みを利用して抽出する場合は関係を抽出できると考えられるが，文頭に文の構造を取りづらい要素があること，*other* などの単語があることから，関係の抽出が行えなくなったものと考えられる．

また，

(9) *BROVANA, as with other beta2-agonists, should be administered with extreme caution to patients being treated with* ***monoamine*** *oxidase inhibitors,* ***tricyclic*** *antidepressants, or*

drugs known to prolong the QTc interval because the action of adrenergic agonists on the cardiovascular system may be potentiated by these agents.

という文は *monoamine* と *tricyclic* の相互作用を示すものではないが，ラベル付きテキストのみでは正しく相互作用なしと出力しているにも関わらず提案手法では誤検出している．前半の “*should be administered with extreme caution to...*” は，多くの学習データがあれば相互作用を示すものとして抽出するようになる可能性が高い一方で，この文ではこれは *other beta2-agonists* との関係であり，この後に登場する薬物同士の相互作用を示すものではないためであると考えられる．

また，学習結果の素性ごとの重さについて，文の中での並びにおいて単語の原型そのものまたはその列についての素性のうち重みの変化が大きかったものを順に並べると表 4.6，4.7 のようになった．ただし，n-gram においては，同様の情報を表す複数の素性の組み合わせが存在することがあるため，ここに入っているからといって必ずしもこの素性の重要度が変化したとは限らない点に注意する必要がある．

ラベルなしデータの追加により関係を示す素性として認識されるようになった項目について着目すると，co-administration of，reuptake などのように比較的登場頻度が低いが重要度の高い単語が入っていることがわかる．co-administration という単語は開発データ中で 21 回しか登場せず，また登場する場合でも相互作用を表さないエンティティペアの場合もあり，データ量を増やしたことによるメリットであると考えられる．

4.2.2 Distant supervision の結果

Distant supervision による手法では，ラベルありデータのみから教師あり学習を行う場合とくらべても著しく低い精度であった．学習に使用できるデータは多いものの精度が低く，正しく学習が行えないためと考えられる．一方で，ランダムに選択されているわけではなく，再現率を下げる方向に調整するとともに適合率が上がっている様子が確認でき，学習手法として意味のある手法であることが確認できる．

4.2.3 知識ベースを片方の view とした co-training

本研究では，知識ベースを片方の view とした co-training による手法においては，教師あり学習を超える結果を出すことができなかった．Co-training では，それぞれの view において一定程度の精度で分類が行えることが必要であるが，現実には知識ベースのみから分類を行った場合の精度が表 3.2 に示されるように低く，また，薬物ペアの組み合わせそのものに関するデータを入力した場合でも，ラベル付けにおける適合率を優先して調整を行った場合でも，40%を超えるデータを得ることができなかった．開発データに 5 回以上含まれるいかなるエンティティペアについても，それが同じ文中に登場する時にそれらの相互作用を示している割合は 40%を超えなかったためである．そのため，薬物の情報のみから充分な分類を行えず，片方の view が不十分であったことが原因であると考えられる．

Co-training では単独で分類を行うことができる view を二つ用意する必要があり，薬物間相互作用においては文の情報を用いずに充分な分類を行うことが困難であると考えられるため，文の精度

表 4.6. ラベルなしデータの追加により関係を示す可能性が高い素性になったもの

場所	単語列	教師あり学習での重み	提案手法での重み
前	antidepressant	-0.104	0.741
前	weekly	0.015	0.655
前	co-administration of [最初のエンティティ]	0.336	0.965
中	[他のエンティティ]	0.336	0.965
後	auc	0.305	0.878
後	[後のエンティティ] a	0.123	0.682
前	antipsychotic	0.008	0.545
中	reuptake	-0.047	0.484
前	with	-0.631	-0.117
前	antiarrhythmic	-0.099	0.406
中	[前のエンティティ] or [後のエンティティ]	-0.546	-0.044
後	[他のエンティティ] , phenytoin	0.105	0.603
前	interaction	-0.311	0.171
前	neuroleptic	0.082	0.560
後	%	0.104	0.583
中	uptake	-0.011	0.467
前	drug	-0.296	0.181
前	concurrent	0.063	0.534
前	[他のエンティティ]	-0.618	-0.147
後	with	-0.604	-0.147

表 4.7. ラベルなしデータの追加により関係を示す可能性が低い素性になったもの

場所	単語列	教師あり学習での重み	提案手法での重み
中	[最初のエンティティ] or other	-0.221	-0.880
後	[後のエンティティ], phenytoin	-0.098	-0.662
中	:	-0.813	-1.364
中	[最初のエンティティ],	-0.241	-0.749
中	[最初のエンティティ], [後のエンティティ]	-0.346	-0.813
前	opioid	0.032	-0.411
中	with	0.555	0.118
前	anesthesia	0.000	-0.421
中	concomitant administration of	-0.307	-0.726
中	contain	-0.032	-0.450
前	(	0.042	-0.373
前	[他のエンティティ],	0.042	-0.373
後	a	0.475	0.065
後	[後のエンティティ] with [他のエンティティ]	-0.186	-0.587
後	)	-0.317	0.717
中	[前のエンティティ] inhibit [後のエンティティ]	-0.052	-0.448
前	interaction study	-0.174	-0.566
中	[前のエンティティ] and	0.564	0.178
前	of [他のエンティティ] with	-0.172	-0.556

の改善のために co-training を適用することは難しいと考えられる．ただし，知識ベースを利用した co-training は困難であると考えられるが，知識ベースを用いず文からの情報での co-training を行うことの有用性を否定するものではない．また，ラベル伝搬アルゴリズムによる半教師あり学習の有用性については [21] で示されている．

なお，この結果は薬物間相互作用については言えると考えられるが，関係抽出一般について成立しないとは考えづらい．関係抽出においては，蛋白質間相互作用の抽出やニュースコーパスからの関係抽出においては distant supervision によっても薬物間相互作用の抽出より精度の高い結果が報告されており [9, 22]，その精度が co-training によってラベルを付すに充分な精度を有しているのであれば，精度改善に役立てると考えられるからである．

4.3 知識ベースをフィルタとして用いた co-training に関する実験

Co-training において，知識ベースのみから付したラベルは精度が低く，そのままでは学習に使用できないと考えられるため，co-training における出力の精度向上のために知識ベースを利用する手法に関する実験も行った．この実験は，学習・テストに使用したデータがこれまでに述べた実験より少ない状態で実験しており，またそのため同一のテストデータとして比較できるわけではないが，この実験セット同士での比較を行うことはできる．

この手法では，co-training の二つの view として，どちらも文からの情報のみを用い，次のように分割した．

- view 1
  - 依存木上での 2 つのエンティティの間のパス上での 1-3 gram
  - 2 つのエンティティの間にある単語の並びについての 1-3 gram
- view 2
  - 2 つのエンティティの前，または後にある単語の並びについての 1-3 gram

また，実験データとしては開発用データセットの一部を用い，分類器としては平均化パーセプトロンを用いた．

この条件で co-training を行うにあたっては，教師あり学習の結果を利用してラベルなしデータにラベルを付している．これに加えて知識ベースを利用するため，ラベル追加のモデルを変更した．つまり，ラベルを追加するにあたり

- 追加しようとしているラベル
  - ラベルを付されたデータからの学習結果からの情報
- 知識ベースからのラベル
  - 追加しようとしているエンティティペアそのものからの情報

の両者が一致した場合のみラベルを付した．この両者は独立したものであり，より精度の高いラベル付けを期待した．変更後のモデルは図 4.3 のようになった．この，co-training で新たに追加するラベルのうち，知識ベースによる分類結果と合致しないものを除外する手法にて，co-training の結果の向上を試みた．

4.3.1 結果

この条件下において，結果は表 4.8 のようになった．

表 4.8. Co-training における F-score

学習方式	F 値
教師あり	66.9%
知識ベースを用いない co-training	68.9%
知識ベースをフィルタとして用いた co-training	67.0%

Co-training 単体において，ラベル付きデータのみで学習を行う場合とくらべて精度が向上しているが，知識ベースによりラベル付けの精度の向上を図った実験では，却って精度が低下している．これは，追加されるデータの数が減少したことによる影響が大きいと考えられ，co-training を行う場合にこのような手法で追加するラベルを制限することによるラベルの精度向上の影響を上回ったと考えられる．

特に，co-training のそれぞれで付されるラベルは，知識ベースのみから付されるラベルより精度が高いことが原因として考えられる．もともと相対的に精度の高い学習データを，より精度の低いデータでフィルタリングすることは，データ全体からの学習結果向上に寄与しないと考えられる．

4.4 考察

以上の実験結果から，提案手法において，ラベルなしテキストに知識ベースによりラベルを付し，これとラベル付きテキストを組み合わせて使用することで，ラベルなしテキストのみ，ラベル付きテキストのみのいずれの場合と比べても高い精度を示すことを確認した．特に，ラベルなしテキストの精度は非常に低いが，その重みを調節することによって，ラベル付きテキストのみから得られる学習結果と比べて大幅に改善できることを示した．また，たとえラベルの付されていないデータであっても学習に使用するデータを増加させることによって精度を改善できることを示した．ただし，開発データにて精度が最も良くなるように調整した後では，全体に占めるラベルの付されていないデータの割合は，調節後の重みを乗じた状態では，ラベルの数を増加させた場合でもそれほど変化していない．このことから，精度が高いデータが多く用意できない場合は，精度の低いデータを多く用意し，個々のラベル付けが誤っていた場合の影響を極力下げることが望ましいと考えられる．

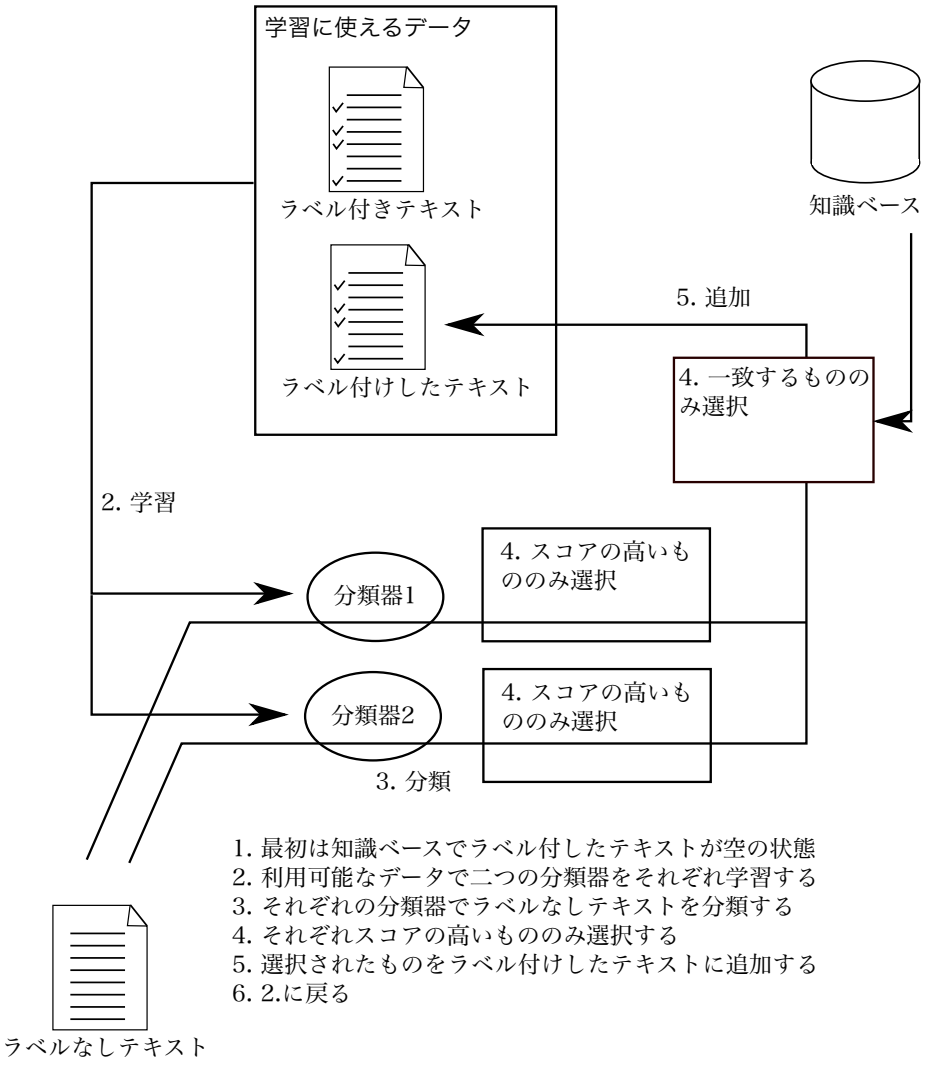


図 4.3. 知識ベースをフィルタとして用いた co-training

また，co-training を利用し，片方の view として知識ベースからの薬物に関する情報のみを利用した場合は，その view の精度が低く，正しいラベル付けが行えないため，学習の精度を改善しないことがわかった．

第5章　おわりに

5.1　まとめ

本研究では，薬物間相互作用抽出において，精度の高いラベル付きテキストに加え，多くのデータが得やすいラベルなしテキストに知識ベースをを利用してラベルを付すことにより得られた精度の低いラベルを使用することで，学習精度を改善する手法を提案し，評価を行い，精度の改善を確認した．

知識ベースを利用して付されたラベルは精度の低く，利用するにあたり，精度の高いデータと同様に扱うことで，学習データの誤りの影響を大きく受けることを防ぐため，ラベル付きデータとくらべて低い重みを付し，ひとつの誤ったデータがある場合の影響をラベル付きデータと比べて小さくすることで，少数のラベル付きデータのみを用いた場合，また，重みを変更しなかった場合と比べても精度が改善することが確認できた．重みを元に計算した場合の全体における自動的に付されたラベルの割合として適した値，元となるラベルなしデータの母集団が共通である本実験においては自動的に付されたラベルによらずほぼ一定であり，多くのデータを用いるほど一つ一つが誤りであった場合の影響を小さくして全体としての傾向を学習できることがわかった．知識ベースなど，テキストそのものによる素性以外から精度の低いラベル付けを行うのに利用可能なデータがある場合や，周辺情報などからある程度のラベル付けが行えるものの，それを co-training の一つの view として扱うには精度が不十分な場合に応用可能であると考えてられる．

また，本研究ではラベルなしテキストの数を変動させて学習を行った結果，ラベルなしテキストを増加させることで精度を改善することができることを確認できた．

そのため，知識ベースを利用してラベルなしデータに精度の低いラベルを付した場合でも，それらに低い重みを付した上でラベル付きデータに追加することが学習精度向上に役立つことが確認できた．

5.2　今後の課題

本研究では Medline を利用して評価を行い，ラベルなしデータを増加させることで精度を改善できることを確認したが，より大きなラベルなしデータを利用できる場合にどの程度精度改善が利用できるかどうかの評価は今後の課題である．本研究では薬物間相互作用での精度改善を試みたが，その他の関係抽出にも応用可能な手法であると考えられる．特に，インターネットを利用してコーパスを集める場合は，薬物間相互作用とくらべてはるかに多くのデータが利用可能であると考えられるが，それらのほとんどはタグ付けされておらず，知識ベースや周辺情報など，不確実な手法を利用

してラベル付けするなどして学習に利用する必要がある．このような場合の応用可能性の評価は今後の課題として挙げられる．

また，生物医学分野においてはまとまった量のラベルの付されたコーパスが利用できるが，学習に必要なラベル付きデータを多く用意することができない場合に，異なる分野のラベル付きデータを学習の元として利用し，それにくわえてラベルなしデータを利用することによる学習について評価を行うことにより，この手法がどの程度ロバストに学習を行うことも，この研究の応用として考えることができる．

参考文献

[1] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pp. 92–100, New York, NY, USA, 1998. Association for Computing Machinery.

[2] Tamara Bobić, Roman Klinger, Philippe Thomas, and Martin Hofmann-Apitius. Improving distantly supervised extraction of drug-drug and protein-protein interactions. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, ROBUS-UNSUP '12, pp. 35–43, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[3] Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. Relation extraction using label propagation based semi-supervised learning. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pp. 129–136, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[4] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 233–240, New York, NY, USA, 2006. Association for Computing Machinery.

[5] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, Vol. 9, pp. 1871–1874, June 2008.

[6] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pp. 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[7] Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pp. 427–434, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[8] Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. Prominer: rule-based protein and gene entity recognition. *BMC bioinformatics*, Vol. 6, No. Suppl 1, p. S14, 2005.

[9] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 541–550, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[10] Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, Christine Jandrasits, Rafael C Jimenez, Jyoti Khadake, Usha Mahadevan, Patrick Masson, Ivo Pedruzzi, Eric Pfeiffenberger, Pablo Porras, Arathi Raghunath, Bernd Roechert, Sandra Orchard, and Henning Hermjakob. The IntAct molecular interaction database in 2012. *Nucleic acids research*, Vol. 40, No. Database issue, pp. D841–6, January 2012.

[11] Ulf Leser and Jörg Hakenberg. What makes a gene name? named entity recognition in the biomedical literature. *Briefings in Bioinformatics*, Vol. 6, No. 4, pp. 357–369, 2005.

[12] Chih-Jen Lin, Ruby C. Weng, and S. Sathiya Keerthi. Trust region newton method for logistic regression. *J. Mach. Learn. Res.*, Vol. 9, pp. 627–650, June 2008.

[13] FaAsma Ben Md. Faisal Mahbub Chowdhury, Alberto Lavelli, and Pierre Zweigenbaum. Two different machine learning techniques for drug-drug interaction extraction. In *Proceedings of the 1st Challenge task on Drug-Drug Interaction*, DDIExtraction2011, pp. 19–26, 2011.

[14] Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the Association for Computational Linguistics: Human Language Technologies*, ACL-HLT '13, pp. 777–782, 2013.

[15] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pp. 1003–1011, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[16] Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. A rich feature vector for protein-protein interaction extraction from multiple corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pp. 121–130, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[17] Kamal Nigam and Rayid Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, CIKM '00, pp. 86–93, New York, NY, USA, 2000. ACM.

[18] Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pp. 113–120, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[19] Foster Provost, Tom Fawcett, and Ron Kohavi. The case against accuracy estimation for comparing induction algorithms. In *In Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 445–453. Morgan Kaufmann, 1997.

[20] Kenji Sagae and Jun'ichi Tsujii. Dependency parsing and domain adaptation with lr models and parser ensembles. In *Proceedings of the Conference on Computational Natural Language Learning 2007*, CoNLL 2007, pp. 1044–1050, 2007.

[21] Min Song, Hwanjo Yu, and Wook-Shin Han. Combining active learning and semi-supervised learning techniques to extract protein interaction sentences. *BMC bioinformatics*, Vol. 12, No. Suppl 12, p. S4, 2011.

[22] Philippe Thomas, Tamara Bobić, Ulf Leser, Martin Hofmann-Apitius, and Roman Klinger. Weakly labeled corpora as silver standard for drug-drug and protein-protein interaction. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining*, BioTxtM2012, 2012.

[23] Philippe Thomas, Illés Solt, Roman Klinger, and Ulf Leser. Learning to extract protein–protein interactions using distant supervision. In *Proceedings of Workshop on Robust Unsupervised and Semisupervised Method in Natural Language Processing*, ROBUS 2011, p. 25, 2011.

[24] David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, Vol. 36, pp. D901–D906, 2008.

[25] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. *J. Mach. Learn. Res.*, Vol. 3, pp. 1083–1106, March 2003.

[26] Zhu Zhang. Weakly-supervised relation classification for information extraction. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, pp. 581–588, New York, NY, USA, 2004. ACM.

[27] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

研究発表報告

- 成川弘樹, 鶴岡慶雅, 近山隆. 知識ベースを利用した半教師あり薬物相互作用抽出. ALAGIN & NLP 若手の会 合同シンポジウム, 2013.
- 成川弘樹, 三輪誠, 鶴岡慶雅, 近山隆. 知識ベースを利用した教師あり薬物間相互作用抽出の改善. 言語処理学会第 20 回年次大会 (NLP2014), 2014. （発表予定）

謝辞

本研究を行うにあたっては，多くの方々にお世話になりました．

指導教員である近山隆教授には，研究の各段階での考え方や方向性についてのご指摘，発表や論文執筆における様々なアドバイスを頂きました．

鶴岡慶雅准教授には，ミーティング等で特に潜在的な問題点や可能性などの観点で鋭いご指摘をいただき，研究にあたって重要なご意見となりました．

また，マンチェスター大学の三輪誠氏，博士の浦晃氏には，研究にあたって壁となる点に関して様々相談に乗っていただくことができ，また多くのご指摘をいただきました．

同期である板持貴之氏，佐藤美沙氏には日頃の発言についても重要な指摘をいただき，特に相談しようと思うことだと自分が認識していなかった点についての指摘をいただけたのは重要なことでした．

個々に名前を挙げた皆様にとどまらず，近山・鶴岡研の皆様には日常的な生活やミーティング，研究の環境において大変お世話になりました．研究を形にするにあたって，非常に重要なものでした．この場をお借りしてお礼申し上げます．