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

成川 弘樹 † 三輪 誠 ‡ 鶴岡 慶雅 † 近山 隆 †
† 東京大学 ‡ マンチェスター大学
{narukawa, tsuruoka, chikayama}@logos.t.u-tokyo.ac.jp,
makoto.miwa@manchester.ac.uk

1 はじめに

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ため，多く用意することが難しい．そのため，正解ラベルを付されていないテキストを多く用意し，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おり [3]，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 [6]．

本稿では，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．この手法の有用性について評価実験を行い，その結果について述べる．

2 関連研究

正解ラベルを付したデータが充分にあれば教師あり学習によって学習を行うことができるが，正解ラベルを付する作業を人間が行うには限界があり，ラベルなしデータと比較して使用できる学習データが限られる．そのため，ラベルなしデータを利用した手法が研究されている [6]．この手法では，実際に報告された関係についての知識ベースを利用することによってラベルなしデータにラベルを付し，ラベル付きテキストと同様に扱うことによって学習を行っている．人の手で学習データにラベルを付す必要がないため多くのデータを利用することができるが，ラベル付きテキストを利用して学習を行った場合と比べると著しく精度が下がっている．

また，関係抽出タスクにおいて，ラベル付きデータとラベルなしデータの両者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 [1] 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エンティティ抽出においてのラベル伝搬アルゴリズムによる半教師学習の有用性が示された．また，蛋白質間の関係抽出において，教師あり学習を行った結果の分類器を用いて高い確度のラベルをつけられるものにラベルを付けることを繰り返すことで，ラベル付きデータを増やしながら学習を行う手法により，半教師あり学習を利用して教師あり学習より高い精度が得られることも示されている [5]．

3 薬物間相互作用抽出

自然言語で書かれた生物医学文献から，薬物間の相互作用を抽出することである．自然言語の文が与えられ，その中の薬物のペアを指定された時に，その文がその薬物間の相互作用を表現しているか否かの分類問題を解くことによって相互作用を抽出することができる．ただし，薬物を指示する語句そのものの抽出は予め行ったものとして考える．

Copyright(C) 2014 The Association for Natural Language Processing.
All Rights Reserved.

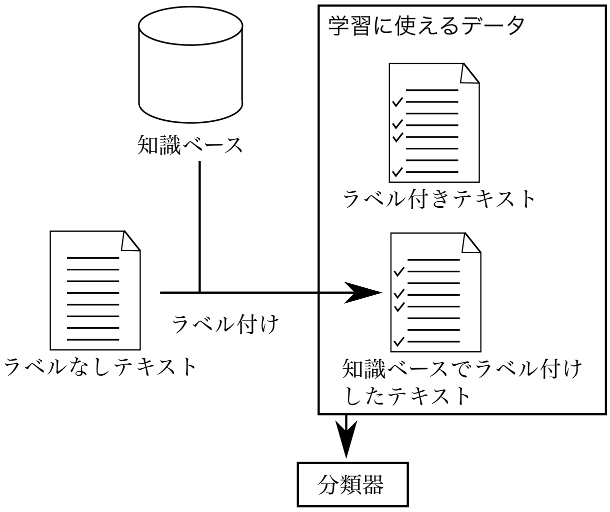


図 1: 学習の流れ

4 提案手法

学習データを利用して分類器を訓練することにより，薬物間相互作用抽出を行う．学習データとして，入力の特徴量とそれに対する正解ラベルの対の群を読み込ませることにより，特徴量から精度よく正解ラベルを推定する分類器を生成するため，学習データを用意する必要がある．

学習に利用するリソースとして，以下のものを使用する．

- ラベル付きテキスト：エンティティの位置と，エンティティ間の相互作用の有無のラベルが付されたテキスト
- ラベルなしテキスト：生テキスト
- 知識ベース：薬物そのもの，また薬物間の相互作用そのものについての情報が入った知識ベース

知識ベースを利用した distant supervision では，学習に使用できるラベルの量は多いものの，そのラベル精度が低く，学習の精度が上がらない問題があった．そのため，正解ラベルのついた学習データを追加し，全てのデータを利用して学習を行う．正解ラベルの付されていない学習データを使用するにあたり，第 4.1 節で述べる手法でラベル付けを行う．

提案手法の，学習の流れは図 1 に示している．

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 [2] の L2 正則化ロジスティック回帰モデルをベースとした．ただし，知識ベースによって付された精度の低い学習データが増えることによって精度が低下することを防ぐ必要がある．そのため，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 (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．ただし，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．

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

4.1 知識ベースの利用

正解ラベルの付されていない学習用テキストデータは，そのままでは分類器にかけることができない．そのため，他の情報を元にラベル付けを行い，学習に利用する．

学習のラベル付けでは， [6] などに従い，以下のモデルを用いている．

- 知識ベースに関係が記された薬物ペアを含む文は，そのペアについての関係に言及している．
- 知識ベースに関係が記されていない薬物ペアを含む文は，そのペアについての関係に言及していない．

5 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．

Copyright(C) 2014 The Association for Natural Language Processing.
All Rights Reserved.

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対応する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 1 のようになった．

表 1: 知識ベースによるラベル付の精度

知識ベースによるラベル	実際に関係があった割合
関係あり	34.3%
関係なし	14.4%
全体	16.6%

6 評価実験

提案手法の有用性を評価するため，以下のそれぞれの手法で学習を行い，その精度を比較した．

- ラベル付きテキストのみを用いた教師あり学習
- Distant supervision を用いた学習
- 提案手法

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．
  - 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
  - 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，大文字小文字の区別なく，-と＝を同一とみなし，空白を無視して一致するものをエンティティとして検出した．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．
- 知識ベース：DrugBank [7] の DDI 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

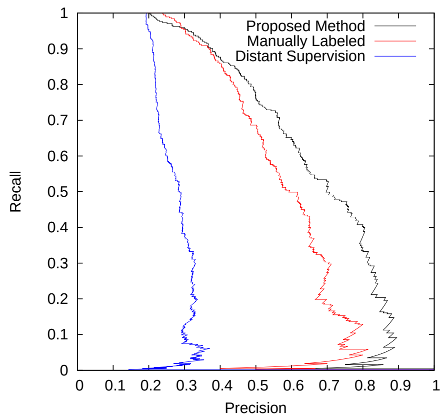


図 2: P-R 曲線

ただし，ラベルなしデータについて，テストデータに含まれるエンティティペアについての情報を知識ベースから参照する必要があるラベル付けは行わず，そのインスタンスは学習データから除外した．

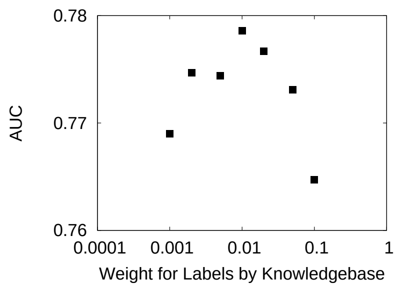
また，テキストに関しては GDep [4] を利用して依存文法の形にパースされたものを利用した．テストデータセットとしては SemEval2013 Task9 用学習データとして公開しているデータのうち，学習データに使用しなかった 1,970 エンティティペアを使用した．

3 つの手法それぞれによる Precision-Recall (P-R) 曲線は図 2 のようになった．P-R 曲線においての Area Under the Curve (AUC) は表 2 のようになった．ただし，ラベル付きのデータを用いた実験では，それぞれ開発用データセットで 3 分割の交差検定を行うことにより，P-R 曲線での AUC が高くなるように式 (1)

表 2: 各手法の P-R 曲線上での AUC

手法	AUC
ラベル付きデータのみ	56.6%
Distant supervision	26.8%
提案手法	65.8%

Copyright(C) 2014 The Association for Natural Language Processing.
All Rights Reserved.



教師あり学習では 0.667 である．

図 3: a を変更した時の AUC(PR) の変化

の C, a を調整した．教師あり学習では $C = 20$，提案手法の実験では $C = 40, a = 0.01$ となった．Distant supervision の実験では $C = 1$ として実験した．

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，図 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．

図 2 に示される通り，distant supervision のみによって多くのデータを元に学習を行っても，教師あり学習を行った場合とくらべて大幅に劣る精度となった．

一方で，提案手法では，ラベルなしテキストに知識ベースでラベルを付したものを利用することで，学習精度を向上する事ができた．特に，関係を示していないものに高い確率で正例であるとする分類を行うケースが減少し，適合率が高い領域において同一の再現率の場合での適合率が向上している．

7 おわりに

本稿では，薬物間相互作用抽出において，ラベルの付されたデータに加えて知識ベースを利用してラベル付けした多くのデータを付することで，人出でラベルの付されたデータのみを用いる場合と比較して高い精度を達成すること，特に，自動で付された低精度なラベルを使用する場合に，低精度であることが予めわかっていればそのラベルの重みを下げることによって学習に利用することができることを示した．

参考文献

[1] Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. Relation extraction using label propagation based semi-supervised learning. In *Proceedings of Coling-ACL 2006*, ACL-44, pp. 129–136, Stroudsburg, PA, USA, 2006. ACL.

[2] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, Vol. 9, pp. 1871–1874, June 2008.

[3] 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 ACL HLT 2011 - Volume 1*, HLT '11, pp. 541–550, Stroudsburg, PA, USA, 2011. ACL.

[4] Kenji Sagae and Jun'ichi Tsujii. Dependency parsing and domain adaptation with lr models and parser ensembles. In *EMNLP-CoNLL*, Vol. 2007, pp. 1044–1050, 2007.

[5] 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.

[6] 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. Proceedings of BioTxtM on LREC, 2012.

[7] 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.

Copyright(C) 2014 The Association for Natural Language Processing.
All Rights Reserved.