Distant Supervision と DB 内情報からの直接学習による薬物間相互作用情報の抽出

西埜 徹　　鶴岡 慶雅
東京大学 工学部電子情報工学科
{nishino, tsuruoka}@logos.t.u-tokyo.ac.jp

1 はじめに

近年、情報抽出や関係抽出の手法は盛んに研究されている分野の一つである。関係抽出の手法により、テキスト中の 2 つのエンティティ間がどのような関係を持つか自動的に判別させることができ、テキストから有用な情報を利用しやすい形で抽出することができる。

関係抽出の応用先の一つとして、薬物間の相互作用を自動的に抽出するタスクがある。膨大な量の医学文献から手動での薬物間の相互作用を表すデータベースの作成は高コストであり、機械的に薬物間相互作用抽出を行う技術は自然言語処理の応用として重要である。

従来の関係抽出の技術では、Zelenko ら [1] の研究に挙げられるように、教師あり学習を用いてエンティティ間の関係の判別を行う手法が主流となっている。しかし、教師あり学習による手法では、学習に十分な量のラベル付き学習データを用意する必要があるという欠点がある。そのため、近年、Mintz らの研究 [2] に代表される、ラベル付き学習データなしでも関係抽出が可能となる Distant Supervision と呼ばれる半教師あり学習の手法が注目されている。

本研究では、Distant Supervision を用いて、ラベル付き学習データなしで薬物間相互作用の抽出を行う際の抽出精度改善を目的とする。分類器の学習を行うにあたって、薬物情報データベース上の分類情報やカテゴリ情報などの知識の援用を行うことで、薬物間相互作用の抽出の精度の改善を行う。知識データベース上の情報の分類への援用にあたっては、素性選択を行うことで、より有効な分類支援を試みる。

2 関連研究

2.1 Distant Supervision

関係抽出でよく用いられている分類手法は教師あり学習による手法である。教師あり学習による関係抽出では、十分な分類精度を得るためには大量のラベル付き学習データが必要となる。しかし、人の手で大量の学習データにラベル付けを行うことはコストが高い。そのため、学習に必要となるラベル付き学習データなし、もしくはわずかなラベル付き学習データから分類器の学習を行う半教師あり学習と呼ばれる手法が注目されている。その一つに、本研究で用いた Distant Supervision と呼ばれる、知識データベースを用いた手法が挙げられる。

Distant Supervision とは、知識データベース上の情報を利用して学習データをラベル付けする半教師あり学習の手法である。Mintz らの研究 [2] では、ラベル付きの学習データなしで分類器の学習を行うにあたって、知識データベース上の関係データを元にしてラベルの付いていない学習データに対して機械的にラベル付けを行い、これを元に分類器の学習を行っている。その研究での Distant Supervision による関係抽出の学習手法は以下の手順で示される。

1. 固有表現抽出器を用いてラベルなし学習データ中のエンティティを抜き出す。
2. 知識データベースを参照して、抽出したエンティティのペアに対して関係の有無をラベル付けする。
3. 教師あり学習の場合と同様にしてラベル付けされた学習データを元にして素性ベクトルを作成する。関係抽出の問題を関係の有無を出力する 2 値分類問題とみなして分類器を学習する。

Distant Supervision によって関係抽出を行う手法の利点としては、ラベル付きの学習データなしでも学習が可能となるため、より多くの分野のテキストに対しての関係抽出の手法の適用が可能となる。一方、欠点として、ラベル付けを機械的に行うため、学習に用いるデータのラベル付け精度は人の手で付けられたものと比較して低くなる点が挙げられる。そのため、教師あり学習に対して関係抽出の精度が高くない。ゆえに、Roth らの論文 [3] で挙げられているように、多

くの研究で関係抽出の精度改善が試みられている。

Distant Supervision による関係抽出の精度を改善する手法として、知識データベース上のエンティティ間の関係情報の他に、エンティティの階層構造を蓄積したオントロジー上の情報を利用する手法がある。Assis らの研究 [4] では、Distant Supervision による関係抽出において、分類器で用いる特徴ベクトルに DBPedia [1] から得られたエンティティ間の階層構造を素性として利用することで、関係抽出の精度を向上させている。

2.2 薬物間相互作用の抽出

薬物間相互作用 (Drug-Drug Interaction, DDI) の抽出は、膨大な量の薬物に関するテキストから相互作用を持つ 2 つの薬物のペアを抽出するタスクである。この薬物間相互作用抽出のタスクは、たんぱく質間相互作用関係の抽出タスク等と並んで、医学分野への関係抽出の応用例の一つである。薬物間相互作用の抽出に関する研究は、DDIExtraction2011 [2] で提供されたデータや、SemEval2013 [3] で提供されたデータなどを元にして行われているが、その多くは Chowdhury らの研究 [5] などで挙げられるように、主に教師あり学習で行われている。

薬物間相互作用を関係抽出する際にも、Distant Supervision の手法を適用する研究も試みられている。例として Bobic らの研究 [6] や Thomas らの研究 [7] などが挙げられる。しかし、これらの Distant Supervision を用いた薬物間相互作用抽出の先行研究では、ラベル付けの精度の低さなどの要因から、いずれも相互作用抽出の精度が F 値にして 30% ～ 40% と低い精度の結果となっている。

2.3 Distant Supervision を薬物間相互作用抽出に適用する際の heuristics な改善手法

これらの薬物間相互作用抽出への Distant Supervision の適用を試みた研究では、低いラベル付け精度を改善し、分類精度を向上させる目的でいくつかのヒューリスティクスを適用している。Thomas らの研究 [8] では、関係抽出を行うにあたって、pos/neg-iword と pos/neg-pair と呼ばれる 2 つの手法を用いている。pos/neg-iword は、正例・負例のそれぞれの学習データのうち、文中に相互作用を表す trigger word が含まれているもののみを学習に使用する手法である。また、pos/neg-pair は、正例・負例のそれぞれの学習データのうち、文中にエンティティのペアがちょうど 1 組含まれているものを選択して学習する手法である。Roller らの研究 [9] では、学習データのうち、2 つのエンティティ間の距離が 5 単語以下の文のみを学習データとして用いる手法 (5w) と、2 つのエンティティ間にコンマが含まれている学習データを無視する手法 (com) を用いて精度改善を試みている。

[1] http://dbpedia.org/
[2] http://labda.inf.uc3m.es/DDIExtraction2011/
[3] http://www.cs.york.ac.uk/semeval-2013/

3 Distant Supervision への薬物データベースの情報の適用

本研究では、Assis らの研究 [4] で提案された、エンティティの階層構造を利用した関係抽出の支援手法を、薬物間相互作用の抽出に適用することを試みる。本研究での提案手法の概要を図 1 で示す。薬物間相互作用の抽出精度改善のために、薬物情報データベース (薬物 DB) 上の相互作用情報以外の情報を援用して分類を行う。薬物間相互作用の抽出に用いる薬品 DB である Drugbank [4] には、相互作用の情報の他にも、薬物の所属するカテゴリの情報や、薬物の分類に関する情報など、薬物に関する種々の情報 (薬物 DB 情報) が収録されている。薬物間の相互作用の起こりやすさは、その薬物ペアの所属するカテゴリや、薬物の分類などの情報との関係があると考えられる。これらの薬物 DB 情報を用いて、関係抽出の分類時に知識データベース上の情報を含めて学習データとして扱うことで、薬物間相互作用抽出の支援を行い、分類精度の改善を行う手法を提案する。

具体的には、分類器に入力するテキストの特徴ベクトルとして、テキスト中から得られる素性に加えて、テキスト中の相互作用を判別するペアとなる薬品に関する分類情報やカテゴリ情報を素性として追加して分類器の学習および分類を行う。

ここで、例として

- There was no evidence of any pharmacokinetic interactions between **ERBITUX** and **irinotecan**.

の文を挙げる。この文中では、薬品 “ERBITUX” と “irinotecan” の間の相互作用は読み取れない。しかし、従来の Distant Supervision による薬物間相互作用の抽出手法では、例文のような相互作用の否定の表現は学習が難しいため、この例は 2 つの薬品間に相互作用があると誤って分類されてしまい、正しい分類が難しいと考えられる。だが、本手法を用いることで、“ERBITUX” が所属する薬物分類情報 (“peptides” 等) と “irinotecan” が所属する薬物分類情報 (“camptothecins” 等) の間では相互作用を引き起こす割合が低いことから、この

[4] http://www.drugbank.ca/

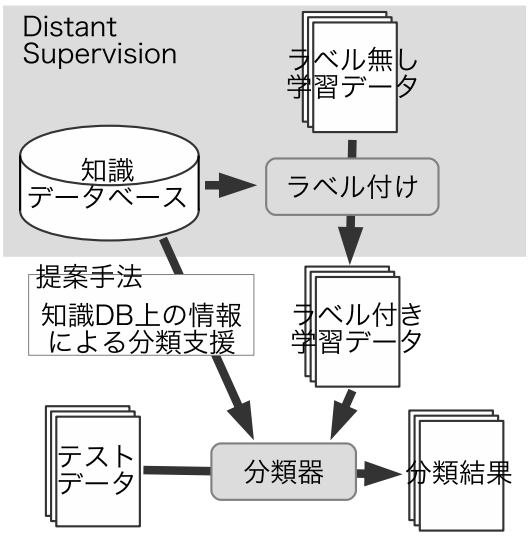

図 1: 薬物 DB の支援による Distant Supervision

2 つの分類情報のペアを素性として利用することで、2 つの薬品間に相互作用はないと正しく分類することが可能になると考えられる。

3.1 素性選択の導入による改善

本研究では、まず素性選択を行わずに薬物 DB 上の情報を関係抽出に用いて予備実験を行った。その際に、使用する薬物 DB 情報のノイズが大きいため、関係抽出の精度が改善されなかった点が問題となった。この問題を解決するために、薬物 DB 上の相互作用に関する情報に基づいて、素性選択を行った。素性選択では、薬物 DB 情報のうち、相互作用を起こしやすい、または起こしにくい情報にあたる素性を選択して使用することを目的とする。

以下の手順で薬物 DB 情報の関係抽出への援用を行った。

1. 薬物 DB 上のカテゴリ情報・分類情報のエンティティ2 つのペアに対し、そのカテゴリおよび分類クラスに所属する薬物ペアのうち、相互作用を引き起こす割合を薬物 DB 上の相互作用に関する情報から計算し、カテゴリ情報・分類情報の素性ごとの相互作用の起こりやすさを求める。
2. 素性選択によって除外される範囲の上限・下限の閾値をパラメータとして与え、1. で計算された相互作用を引き起こす割合がこの範囲内に含まれるカテゴリ情報・分類情報の素性を除外する。
3. 関係を分類する 2 つの薬物がそれぞれ所属するカテゴリ情報・分類情報のペアを、文章中から生成された素性ベクトルに追加する。この素性を用いて分類器 (SVM) の学習・分類を行う。

データ	関係あり	関係なし	合計
学習データ	41,084	425,010	466,094
開発データ	3,788	22,217	26,005
テストデータ	755	6,271	7,026

表 1: データセット中の正例・負例の数

4 実験と結果

3 章で述べた提案手法によって、Distant Supervision による薬物間相互作用の抽出精度が改善されることを確認する実験を行った。

4.1 実験データセット

この実験にあたって使用したデータを以下に示す。

ラベルなしの学習データ PubMed [5] 上で配布されている医学分野の論文を集めた医学文献データベースである MEDLINE のコーパス [6]

知識データベース 薬物 DB である DrugBank 上の情報 [7] のうち、薬物名・薬物間相互作用にまつわる情報、及び薬物の分類情報やカテゴリ情報

開発データ・テストデータ 開発データ・テストデータにはそれぞれ SemEval2013 および DDIExtraction2011 で薬物間相互作用の抽出タスク用として配布された、医学文献から作成されたラベル付きのデータを用いる。SemEval2013 のデータは、学習データとして提供されている分を、今回の実験では開発データとして分類精度などを評価する目的のみで用いる。

これらのデータのうち、相互作用関係があるデータ、および相互作用の関係のないデータの数を集計した結果が表 1 である。この表で示されるように、データセット中の正例・負例の割合には大きな偏りがあるため、学習時には正例・負例の比を 1:1 になるように学習データをランダムで選択して学習に用いた。

4.2 Distant Supervision による関係分類

本実験で関係分類に用いた分類器は、線形 SVM ライブラリである LIBLINEAR [8] である。関係分類に用いた文中の特徴の一覧は以下の 2 つである。

語彙的な特徴 前後の単語・2 エンティティ間にある単語の 1-gram 及び 2-gram

[5] http://www.ncbi.nlm.nih.gov/pubmed

[6] http://www.nlm.nih.gov/で配布されている MEDLINE/PubMed のコーパスのうち、ベースラインとして配布されているデータの 2014 年版

[7] http://www.drugbank.ca/downloads で提供されている Drugbank 上の全薬物の情報のデータベース (2014 年 9 月 8 日更新,ver4.1)

[8] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

手法	適合率	再現率	F 値
薬物 DB 情報なし	39.1	37.7	38.4
薬物 DB 情報あり	**41.4**	30.7	35.3
薬物 DB 情報あり (素性選択; 閾値 0.0004〜0.9)	38.0	41.9	**39.8**
[Thomas et al., 2013] [7]	33.0	**44.1**	37.7

表 2: Distant Supervision による関係抽出結果

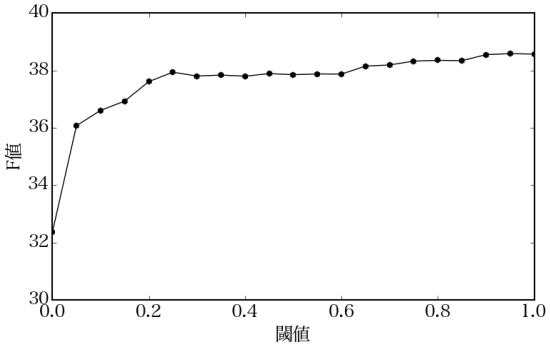

図 2: 素性選択で除外される閾値の上限と F 値の関係

依存木 (文章中の文法構造の依存関係を木構造として表現したもの) 上の 2 つのエンティティを結ぶ最短パス上の単語

文法的な特徴 前後の単語・2 エンティティ間にある品詞の 1-gram 及び 2-gram

依存木上の 2 つのエンティティを結ぶ最短パス上の品詞

この 2 つの特徴に加えて、3 章で述べたように、薬物 DB 上の情報を特徴に加えて関係分類に用いた。

また、関係抽出の精度改善のために、2.3 節で述べたヒューリスティクスによる改善手法のうち、pos/neg-iword と pos/neg-pair、及び 5w と com を適用した。5w に関しては、開発データを用いて、文章中の 2 つのエンティティ間の単語距離の閾値のパラメータを調整した結果より、8 単語を閾値と決定した。

4.3 実験結果

この提案手法を用いて分類した結果を表 2 に示す。薬物 DB 上の情報を素性選択なしで利用した時は、薬物 DB 情報を利用しない時と比較して F 値が低下した。一方、素性選択を適用した時、分類精度は、薬物 DB 情報を利用しない時と比べて改善された。この事から、薬物 DB 情報は薬物間相互作用の精度改善に有効であると確認された。

また、開発データで素性選択を用いて関係抽出を行った際の素性選択の閾値と分類精度の関係を図 2・図 3 で示す。この結果から、素性選択するときの素性が除外される範囲は 0.0004 と 0.9 の間と決定した。

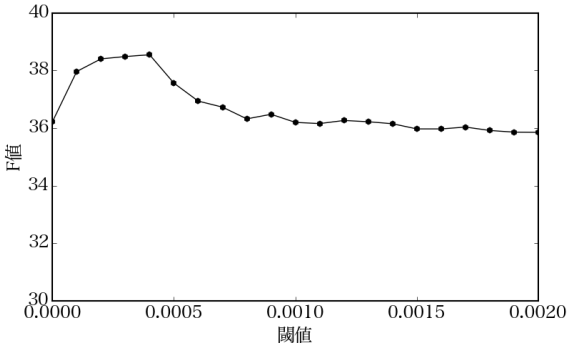

図 3: 素性選択で除外される閾値の下限と F 値の関係

5 おわりに

本研究では、薬物間相互作用の抽出に Distant Supervision を適用する際、精度改善のために薬物 DB 情報を援用して分類を行う手法を提案した。薬物 DB 情報を用いる際に、素性選択を行い、薬物 DB 情報のうちノイズとなっているものを取り除くことで、精度改善されることが確認された。

本研究では、薬物 DB 情報を Distant Supervision の分類に用いたが、4.3 で述べたように、素性選択するときの閾値は 0.0004 と 0.9 と決定した。しかし、この時、相互作用を引き起こす割合の高い素性が殆ど素性選択によって除外されている点など、必ずしも最適な素性選択が行われているとは言い難い。そのため、より優れた素性選択を行う手法を検討することが今後の課題である考える。

参考文献

[1] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. In *ACL*, 2002.

[2] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *ACL-IJCNLP*, 2009.

[3] Benjamin Roth, Tassilo Barth, Michael Wiegand, and Dietrich Klakow. A survey of noise reduction methods for distant supervision. In *AKBC*, 2013.

[4] Pedro HR Assis and Marco A Casanova. Distant supervision for relation extraction using ontology class hierarchy-based features. In *ESWC*, 2014.

[5] Md Faisal Mahbub Chowdhury and Alberto Lavelli. FBK-irst: A Multi-Phase Kernel Based Approach for Drug-Drug Interaction Detection and Classification that Exploits Linguistic Information. In *Semeval*, 2013.

[6] Tamara Bobić, Roman Klinger, Philippe Thomas, and Martin Hofmann-Apitius. Improving distantly supervised extraction of drug-drug and protein-protein interactions. In *ROBUS-UNSUP*, 2012.

[7] 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 *BioTxtM*, 2012.

[8] Philippe Thomas, Illés Solt, Roman Klinger, and Ulf Leser. Learning protein protein interaction extraction using distant supervision. In *ROBUS*, 2011.

[9] Roland Roller and Mark Stevenson. Applying UMLS for Distantly Supervised Relation Detection. In *EACL*, 2014.