Wikipediaを用いた遠距離教師あり学習による専門用語抽出

宮崎 亮輔 * [1]　小町 守 [1]　疋田 敏朗 [2]　柏倉 俊樹 [2]

[1] 首都大学東京　[2] 株式会社トヨタIT開発センター

1　はじめに

専門用語抽出はコーパスから専門用語を抽出する技術である．専門用語のような重要な用語を辞書として保持しておくことは，文書分類や情報検索などの自然言語処理技術を用いたアプリケーションにおいて重要である．

従来，専門用語の抽出は専門家の人手によらねばならず，大量な人手と時間がかかる作業であった．そのため常に更新された辞書を保持することは困難であった．そこで，コーパスから自動で用語を抽出する手法が研究されている．一つの方法としてブートストラッピング法という手法がある．これは，人手で作成した少数のシード辞書をもとに繰り返しコーパスから用語を抽出する方法である．しかし，ブートストラッピング法を利用する上では意味ドリフトの問題が存在する．他にも，コーパスにアノテーションをして教師あり学習を行うことで専門用語を抽出する方法も考えられるが，アノテーションのコストがかかってしまう．

そこで，本論文では**遠距離教師あり学習（distant supervision）**を用いてWikipediaから得たシードをもとにコーパスに自動でアノテーションすることで専門用語を抽出する方法を提案する．ブートストラッピング法と比べて非常に多くのシードを用いるが，Wikipediaから自動でシードを獲得することでシードを用意する手間を軽減した．また，実験により84 %の適合率で専門用語を抽出できることを示した．

2　関連研究

2.1　ブートストラッピング法

ブートストラッピング法は自然言語処理における情報抽出の一般的なフレームワークである [2,4]．ブートストラッピング法は獲得対象となるクラス（例:is-a関係）のインスタンス（例:(cat, animal)）をシードとして与え，コーパスからインスタンスと共起するパターンを抽出し，抽出した共起パターンを用いて新たなインスタンスを抽出する．この手順を反復的に繰り返し，少数のシードインスタンスから大規模なインスタンスの集合を再帰的に獲得する手法である．

このブートストラッピング法では，反復処理を繰り返していくうちにシードインスタンスと関係のないインスタンスを抽出してしまう問題が知られており，意味ドリフトと呼ばれている．あらかじめ手元に多くの用語（シードデータ）がありそれに加えて更に新しい用語を取得するような場合には，反復処理をする必要がないため，提案する手法では意味ドリフトの問題を考慮する必要なく新たな専門用語を抽出することができると考えられる．

2.2　遠距離教師あり学習

これまでに関係抽出のタスクで遠距離教師あり学習が成果をあげている [1,3]．遠距離教師あり学習では教師あり学習や半教師あり学習のようにラベルつきのコーパスを必要としない．その代わり遠距離教師あり学習では，大量の知識ベースと大量のラベルなしコーパスを用いる．

Mintzらの遠距離教師あり学習を用いた手法では，関係ペアとその関係を表す3つ組をFreebase[1]から大量に抽出し，それを大量の知識ベースとした．また，Wikipediaのダンプデータを大量のラベルなしコーパスとして利用した．この大量のラベルなしコーパスに対して，知識ベースに保持しておいた3つ組にマッチする文があれば正例だと見なして学習する．

例として，知識ベースには “(Obama, Hawaii, Live in)”,“(Obama, Hawaii, Born in)” を含む関係を表す3つ組を多数保持しているとする．つまり，“Obama” と “Hawaii” の組に対して “Live in” と “Born in” という2つの関係を知識ベース内に持っていることになる．ここで，“Obama was born in Hawaii.” という文がラベルなしコーパスに出現した場合を考える．この

*miyazaki-ryosuke@ed.tmu.ac.jp

[1] https://www.freebase.com

“Obama was born in Hawaii.” という文の学習を行うが，正解ラベルが “Born in” であると見なして関係分類器の学習を行うだけでなく，加えて正解ラベルが “Live in” であると見なして学習も行う．すなわち，トレーニングデータ内の文に対して，保持している知識ベースとマッチするペアが存在すれば，そのペアに対して保持している知識ベース内のすべての関係タイプを正解ラベルと見なして関係分類器の学習を行う．

上記の例での “Live in” を正解ラベルだと思って学習を行う例のように，間違ったラベルを正解だと思って学習してしまうこともある．しかし，教師あり学習のような少ないリソースだけでなく，大量のラベルなしコーパスを用いることが可能なため，素性の表現がより豊富になるという利点がある．そのため，いくらかの雑音があったとしても小さいコーパスの教師あり学習と比べても性能を向上することが可能になった．

3 遠距離教師あり学習による専門用語抽出

本論文では，ある特定の専門分野で通用される語彙をその分野の専門用語とし，その分野に適応した専門用語の抽出方法を提案する．

上に述べたように，これまでに関係抽出のタスクで遠距離教師あり学習の手法が成果をあげてきた．関係抽出のタスクでは，2 対の単語ペアからその単語間の関係を限られた選択肢の中から選ぶという分類問題を遠距離教師あり学習で解いていたが，関係抽出と違い用語抽出では用語の区切りを考慮して分類する必要がある．本論文では形態素区切りが専門用語の区切りとなると仮定し，複数形態素で一つの専門用語が構成されることがあるので，系列ラベリング問題を遠距離教師あり学習で解き，その結果から専門用語を抽出する．提案する手法については以下に詳しく説明する．

1. シードとなる用語を用意する
2. トレーニング用のラベルなしコーパスを用意する
3. シード内の用語をもとに，ラベルなしコーパスに対して自動でラベルを付与する
4. 自動で付与したラベル付きコーパスをもとに，系列ラベリングによる遠距離教師あり学習を行う
5. 学習したモデルを用いて抽出対象のコーパスを解析する
6. 解析結果から新たに得られた用語を専門用語として抽出する

1 ステップ目では遠距離教師あり学習にて必要になる大量の知識ベースを準備する．本論文では Wikipedia 内のあるカテゴリに属するタイトルを全て抽出し，そのタイトルをそのカテゴリの専門用語だと仮定している．もちろん専門用語でないタイトルも含まれるが，ここでは無視する．

2 ステップ目では遠距離教師あり学習にて必要になる大量のラベルなしコーパスを準備する．本論文では Wikipedia から抽出して利用している．

3 ステップ目で，遠距離教師あり学習を行うために擬似的なラベル付きコーパスを作成する．ラベルなしコーパスに対してはあらかじめ形態素解析を施しておく．例えば，知識ベース内に “シフトレバー” という用語があった場合を考える．ラベルなしコーパス内に “シフト—レバー—を—動かす” のように知識ベースと一致する単語が現れると，単純なパターンマッチによってその文に BIO のラベルを付ける．この場合は，“シフト (B)—レバー (I)—を (O)—動かす (O)” というラベルが付与される．

4 ステップ目で遠距離教師あり学習を行う．3 ステップ目にて擬似的に作成されたラベル付きコーパスをもとに，教師あり学習と同様に系列ラベリング問題を学習することができる．

5 ステップ目で対象とするコーパスには，抽出する専門用語が多く含まれると予想できるコーパスや大規模なデータを想定する．例えば，Wikipedia の全データなどが考えられる．この対象とするコーパスに対して 4 ステップ目で学習したモデルを用いて解析を行う．すなわち自動で対象とするコーパスに BIO の系列ラベルを付与する．

6 ステップ目によって新たな専門用語を抽出することができる．5 ステップ目で解析された結果から一度でも BI のラベルのついた用語を抽出するのである．その中からもともと知識ベースに存在していた用語を差し引いた残りが新たに抽出された専門用語となる．

これらのステップを踏むことで，人手によるアノテーションやデータベース構築なしに Wikipedia からシードデータベースを抽出し遠距離教師あり学習を行うことで新たに専門用語を抽出することが可能になる．

4 専門用語抽出とその妥当性の実験

遠距離教師あり学習による専門用語抽出の実験を行う．本実験では 2 種類の実験を行う，1 つ目は，専門用語を抽出できるかの妥当性の実験を行う．これには Wikipedia の記事データ（2014 年 12 月時点）を利用して交差検定を行う．2 つ目は実際に新たな専門用語

表 1: Wikipedia から抽出したカテゴリ名とタイトル名の例

カテゴリ名	タイトル名
自動車工学	アクティブスタビリティコントロール, アクティブ・ヨー・コントロール, アダプティブ・フロントライティング・システム, 外輪差, 過給圧
自動車のエンジン	PRV エンジン, ノーススターエンジン, ポルシェのエンジン一覧, ユニフロー掃気ディーゼルエンジン, リアエンジン
ブレーキ	エアブレーキ, ジャダー, サーボブレーキ, ディスクブレーキ, ブレーキ, 圧縮開放ブレーキ, 自動空気ブレーキ, 自動ブレーキ, 真空ブレーキ

表 2: 学習に利用した素性テンプレート

表層形に関する素性	$w_{-2}, w_{-1}, w_0, w_1, w_2$ $w_{-1}w_0, w_0w_1$
文字種に関する素性	$t_{-2}, t_{-1}, t_0, t_1, t_2$ $t_{-2}t_{-1}, t_{-1}t_0, t_0t_1, t_1t_2$ $t_{-2}t_{-1}t_0, t_{-1}t_0t_1, t_0t_1t_2$
組み合わせ素性	w_0t_0

を抽出する実験を行う．これには国土交通省の自動車リコール不具合情報データベースを利用する．

4.1 共通する実験設定

いずれの実験も専門用語のドメインとして自動車の専門用語を対象とする．以下に共通する実験設定を記述する．各ステップ番号は 3 節でのステップ番号と対応している．

1. Wikipedia から “自動車工学” カテゴリに属する記事タイトルを抽出する
2. 抽出したタイトルの記事本文を取得して形態素解析をする
3. 抽出したタイトルをシードデータベースとして，形態素解析済みコーパスに対して系列ラベルを自動で付与する

1 ステップ目の記事タイトルの抽出はカテゴリ名をもとに行う．対象ドメインを自動車の専門用語としたので，自動車関連の記事が多く含まれるであろう “自動車工学” カテゴリをルートカテゴリに設定した．更に自動車工学カテゴリ以下の包含関係にあるカテゴリからもタイトル名の抽出を行う．すなわち，自動車工学カテゴリに含まれるタイトルと自動車工学カテゴリの子カテゴリに含まれるタイトルを抽出した．これにより，最終的に 690 の記事タイトルが得られた．得られたカテゴリ，タイトルの例を表 1 に示す．

2 ステップ目では，これらのタイトル名を含むラベルなしコーパスを取得するために，同タイトルの記事本文を同じく Wikipedia から取得する．結果，全 70,921 文のラベルなしコーパスを取得した．また前処理として得られた記事本文データに対して形態素解析を行う [2]．このとき，できるだけ短単位に区切って系列ラベルを付与することでより細かい粒度の素性情報を得られるため，形態素解析器 MeCab 0.996（辞書は UniDic 1.3.12）を用いた．

3 ステップ目では遠距離教師あり学習を行うために，ラベルなしコーパスに対して自動でラベルを付与する．シードデータベース内に存在するあるタイトル名と単純にマッチする形態素列がラベルなしコーパス内に存在する場合，そこに系列ラベルを自動で付与する [3]．シードデータベース内に存在するタイトル名すべてに対して同様に系列ラベルの付与を行う．

以上のステップによって，擬似的にラベル付きコーパスを作成する．以降では，この擬似的なラベル付きコーパスを用いて実験を行う．

学習の素性には，形態素の表層形の素性と，その形態素がカタカナだけで構成されているか，英字だけで構成されているか，それ以外かの 3 値の素性の 2 種類を用いた．前者を w，後者を t と表して，実験に利用した素性テンプレートを表 2 に示す．また，系列ラベリング問題の学習および解析には CRF++ 0.58 を利用した．

表 4: Wikipedia を用いた妥当性実験の結果

	適合率	再現率	F 値
平均値	69.45%	39.27%	49.80

4.2 妥当性の実験

実験設定 本実験では遠距離教師あり学習によって専門用語の抽出をすることができるかどうかの妥当性を評価する．4.1 節で作成したラベル付きコーパスを 10 分割して [4] 交差検定を行い，テストデータ内に含まれる専門用語をどれだけ当てられるかを実験する．

実験結果 4.2 節の実験の結果を表 4 に示す．表 4 は 10 分割交差検定の結果を表している．再現率は約 40

[2] 形態素単位ではなく文字単位に区切っても動作するが，本実験では形態素単位を扱う．

[3] 複数形態素に対して系列ラベルを付与する．形態素区切り以外でマッチした場合にはラベルの付与はしない．

[4] このとき，各トレーニング時にはテストデータ内に出現する各正例（すなわち系列ラベルが付与された各タイトル名）をトレーニングしない．

表 3: 国土交通省の自動車のリコール不具合コーパスからの専門用語抽出実験の結果得られた単語

	分類	件数	例
専門用語（84 件）	部品	64 件	グローランプ, ドアバイザー, ディスクパッド, スロットルペダル, エアーバッグ, リアホイールシリンダー
	システム	14 件	サスペンションコントロールシステム, グローコントロール, 2 輪駆動, アクティブステアリング
	現象，動作	6 件	ノッキング現象, ジャダー現象, ランオン現象, レストブレーキ, 尾灯点灯, ノーブレーキ
その他の単語（16 件）	オノマトペ	4 件	ガクガク, ガタガタ, コトコト, コツコツ
	A の B，A な B	8 件	キックダウン後の振動, ブレーキペダルからの振動, 前照灯の整備, タイヤの振動, 走行の振動, 微細な振動
	その他の（一般的な）名詞	3 件	ドアブレーキ, 温度計, 電気消費率
	区切り間違い	1 件	尾灯・制動（本来は “尾灯・制動灯”）

%と低いが，適合率が約 70 %で新たな専門用語を抽出できていることを確認できた．

しかし，再現率が低い割に，適合率も高いとは言えない．これは専門用語抽出と言うよりもタイトル推定になってしまっているためだと推測できる．すなわち，新たに抽出した専門用語が実際には専門用語であるにも関わらず，シードデータベースに存在しなかったために負例になってしまっているというケースが考えられる．エラー分析を行った結果，交差検定のある分割において False Positive 143 件中 86 件（約 60%）が上記ケースに該当する例であった．この分割においては適合率が真には約 81.4%であったはずだが，53.4%まで下がってしまっていた．

4.3 専門用語抽出の実験

実験設定 本実験では，実際に専門用語の抽出を試みた．学習には 4.1 節で作成したラベルの付与されたコーパスを用いる．本実験では交差検定のように分割は行わずコーパスすべてを学習に用いる．解析対象には国土交通省の自動車のリコール不具合情報データベース [5] をクロールして取得したコーパスを用いる．自動車の不具合やリコールに関する情報の文を約 30,000 文集めることができた．このリコール不具合コーパスを解析し，そこから専門用語抽出を行う．

実験結果 リコール不具合データを解析した結果，新たに 100 件の用語を抽出した．抽出した用語の中には自動車の専門用語と言えない単語も存在したが．抽出した用語を人手で確認すると適合率は 84 %であった．抽出した用語の一部を表 3 に示す．

表 3 に見られるように，抽出して得られた新たな用語の中にも専門用語ではない単語も含まれていた．その他の単語として抽出された用語の中にはオノマトペや名詞句になっているものが多い．人手による作業が必要になってしまうが，抽出して得られた単語に対してルールベースなどのフィルター（例えば “A の B” のパターンを除外する等）をかける操作を行ったり，もともとのシードデータベースのクリーニングを行ったりすることで，適合率をさらに上げることができるのではないかと考えられる．

5 おわりに

本論文では遠距離教師あり学習による専門用語抽出を可能にした．ブートストラッピング法と比べて多くのシードデータを必要とするが，Wikipedia から自動で取得することで人手の作業を減らしている．

参考文献

[1] Gabor Angeli, Julie Tibshirani, Jean Y. Wu, and Christopher D. Manning. Combining distant and partial supervision for relation extraction. In *EMNLP*, pp. 1556–1567, 2014.

[2] Marti A. Hearst. Automatic acquisition of hyponyms. In *COLING*, pp. 539–545, 1992.

[3] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *ACL-IJCNLP*, pp. 1003–1011, 2009.

[4] Michael Thelen and Ellen Riloff. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *EMNLP*, pp. 214–221, 2002.

[5] http://carinf.mlit.go.jp/jidosha/carinf/opn/index.html