Web 文書を利用した半教師あり用語抽出*

近藤 光正　乾 健太郎　松本 裕治
奈良先端科学技術大学院大学
{mitsu-ko,inui,matsu}@is.naist.jp

1 はじめに

近年，時事的な Web 文書から，製品の評判，機能など，必要な情報だけを自動抽出する情報抽出技術が注目を集めている．しかしながら，情報抽出の対象となる分野は幅が広く，現在用いられている主な情報抽出技術では，それぞれの分野に特化した用語辞書を人手で構築する必要があり，用語辞書の規模は情報抽出システムの精度を大きく左右する．

そこで本研究では，できるだけ少ない人手コストで大規模な用語辞書を構築することを目標とし，複数の特定の分野に特化した用語辞書をより簡単に構築することを目指す．

2 関連研究

近年，少量の訓練データと未知のアンラベルドデータを学習の一部として用いることで，より優れた学習を実現する半教師あり学習手法が注目を集めている．アンラベルドデータを用いた半教師あり学習手法の代表的な手法が Blum らが提案した Co-Training [1] である．後に Co-Training は多視点学習（Multi-View Learning）の一手法として位置付けされるようになるが，これらの学習の基本的な考え方は，異なった視点を持つ分類器を複数個作成し，お互いの分類器の持つ情報を交互に補間し合うことでより良い学習を目指すものである．彼らの手法は，最初に，分類対象の文書が持つ単語の素性と，ハイパーリンク先の文書が持つ単語の素性にそれぞれ素性を分割することによって，二つの分類器を作成する．そして，各分類器が出力した確信度の高い文書の一部を互いに相手の訓練データに追加し，再度それぞれの分類器の学習を繰り返すことで，素性を分割しない従来の手法よりも優れた学習を実現した．

Nigam らは Co-Training と EM アルゴリズムを組み合わせた Co-EM [5] [6] の提案と共に，Blum らが Co-Training が実行できる条件として挙げた内容についての実験を行っている．そして，素性が完全に独立するテストデータを人工的に作成することで，Co-Training に理想的な環境で実験した結果と，ランダムに素性を分割して実験を行った結果を比較し，お互いの素性が独立であればあるほど優れた学習が可能であることを実証した．

Collins [7] は Yarowski が多義性解消に用いた手法 [2] と Co-Training を組み合わせることで，少量の訓練データから大量の英語固有表現を高精度に分類することに成功した．Co-Training を実行するのに必要な条件である各素性集合の分割は，分類対象の固有表現そのものから得られる素性（用語内素性）と，固有表現の周辺部にある素性（用語外素性）に分割した．

他にも，日本語固有表現抽出にブートストラップを適用した論文 [11] や医学分野の用語抽出に Co-Training を適用した論文 [12] がある．これらの論文からは，言語特有の用語外素性の問題や，抽出対象の用語によっては用語外素性の効果が異なるということがわかる．

また，用語外素性はテストデータ中の用語の出現頻度によって精度が異なってくることも考えられる．すなわち，事例を多数持つ用語に関しては，異なった表記の用語間で同じ用語外素性をもつ可能性が高いので高い精度が期待できるが，用語の事例が少ない場合にはこの可能性に関して期待が持てないということである．実験により検証したわけではないが，Collins の手法は事前に用意できる文書量が少ない場合には，効率的な抽出ができない可能性がある．

Co-Training を実現するのための条件として素性の分割が必要であるが，各素性から学習した分類器の片方の精度が Co-Training の各ラウンドで支障をきたす場合，Co-Training の実現はできない．Muslea2002[4] や Denis2003[3] に記載があるように，Co-Training は，実行条件が非常に厳しい手法であるため，抽出したい用語の特性や収集できる文書量の違いによって手法の正否が異ならないような頑健な手法が望まれる．

3 提案手法

本稿では比較的小規模な専門文書集合と少量の用語辞書が事前に与えられている状況について考える．そして，Web から用語事例を多数獲得し，さらに用語事例の統計的な分布の確率モデルを素性とすることで頑健な半教師あり用語抽出を実現する手法を提案する．

3.1 用語抽出の流れ

まず初めに，事前に与えられている初期文書集合の形態素解析および，係受け解析を行う．次に，先ほどの解析結果から，名詞，記号，未知語等の連続語を用語候補として抽出し，用語候補が事前に用意した用語辞書（訓練辞書）に含まれている場合にその用語候補を訓練データとし，用語辞書に含まれない残りの用語候補をテストデータとする．そして，機械学習器[1] によって学習したモデルから，テストデータを解析する．用

*Semi-Supervised Term Extraction Using Web Document
Mitsumasa Kondo, Kentaro Inui, and Yuji Matsumoto
Nara Institute of Science and Technology

[1] Support Vector Machines [10] を用いる

語には様々な出現パターンがあり，それらすべての事例を評価関数を用いて用語か否かを判別する．ここまでが，本稿における教師あり用語抽出の主な流れである．さらに半教師あり学習を用いて用語抽出を行う場合には，先ほど出力した確信度の高い用語候補の一部を訓練辞書に追加し，そして再度学習を行う作業を繰り返し実行することでより優れた学習を実現する．図1に本稿における用語抽出の主な流れを掲載する．

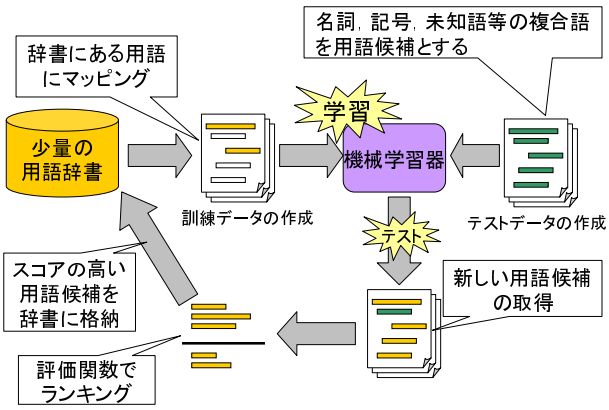

図 1: 用語抽出の主な流れ

3.2 Web からの用語事例の取得

初期の文書集合だけでは用語判定のための事例が少ない場合がある．そこで本稿では，初期文書集合から抽出したすべての用語候補を Web サーチエンジン[2] に自動入力することでより多くの用語の事例を獲得する．取得は Web 文書ごとに行うが，Web 文書集合の事例を解析する際には初期文書集合から抽出した用語候補のみを対象とする．Web を用いることで，過去の新聞コーパス等にはない，製品名等の時事的な要素を含む用語候補の事例を集めることができるメリットもある．

3.3 用語性判定のための評価関数

本稿では SVM が算出した決定関数 (分離平面からの距離) を事例のスコアとするが，用語判定のためには，全ての事例から用語の代表スコア（確信度）を算出する必要がある．様々な評価関数を用いて実験した結果，最も正例側にある事例を用語の確信度とする手法を採用する．

3.4 学習に使用する素性

3.4.1 用語内素性

用語を構成する形態素列 用語を構成する形態素列の表記，品詞，文字種の素性で，用語内における位置を考慮して入力する．

CaboCha の固有表現解析結果 CaboCha が出力する固有表現解析結果を素性として加える．なお，非固有表現を表す O タグは除外する．

用語の形 用語の文字単位での文字種の解析結果を素性として加える．また，同じ文字種が連続した場合を考慮した素性と考慮しない素性の 2 種類を素性とする．

[2]本稿では Google を使用した．

用語内に含まれる記号，記号列の並び 用語に含まれる記号を素性とし，さらに記号列の並びを素性として加える．

品詞列の並び 品詞列の並び方から専門用語を抽出しようという研究があり，一定の成果を納めている．同じ品詞が連続した場合を考慮した素性と考慮しない素性の 2 種類を素性として加える．

用語の長さ 文字列や形態素列が長ければ長いほど専門用語である確率が高いという報告がある．3.3 節で述べたように SVM で学習した分離平面との距離をスコア関数としているため，0〜1 の値になるように正規化を行った．

文書集合での用語の出現分布 専門用語には特徴的な出現分布がある．そこで，統計的な尺度である，tf，df，idf，tf·idf，,df2，df2/df, また，Google のヒット数から df と idf を算出し，各値を 0〜1 に正規化して使用する．

3.4.2 用語外素性

用語の周辺文字列 未知の用語を説明する際に，鉤括弧で括ったり，用語の省略形を括弧で括る場合が良くある．用語周辺部の前後±2 の形態素の位置を考慮した素性として加えた．

用語の係先・係元の主辞 専門用語抽出において，係先と係元は用語の使われ方や用語のクラス性に大きな影響を持つ．そこで，係先と係元の主辞のみを素性とする．

分類語彙表による拡張 係先 · 係元の主辞を分類語彙表 [13] から辞書引きし，全ての階層の構成をそれぞれ素性とする．

PLSI を用いた用語の周辺部の統計的な素性 これまで述べてきた用語外素性は，用語内素性と比較すると効果が大変低い．本稿では，用語辞書構築を目的としているため，各用語の出現事例一つ一つを正確に抽出するのではなく，各用語のすべての出現事例から用語性を判別すれば良い．また，Web サーチエンジンを用いて用語の事例数を大幅に増やすことに成功しているので，用語の周辺部の素性を統計的に扱うことによる用語のクラス判別は大変効果的であると考える．そこで，全ての出現事例の用語外素性ベクトルの和を用語毎に算出し，それらのベクトルを f，用語を t，隠れクラスを z とし，Probabilistic Latent Semantic Indexing(PLSI) [8] を実行する．そして，PLSI の出力した $P(z \mid t)$ を用語外素性とする．PLSI を実行する理由として，用語外素性ベクトルの和を正規化し，さらにスムージングを行うことでクラス判定のための素性として有効であると考えたからである．以下に，本稿で使用した PLSI の式を掲載する．

$$P(f,t) = \sum_{i=1}^{n} P(f \mid z_i) P(z_i) P(t \mid z_i)$$

Collins の手法は，一部の定型的な用語用例から用語を抽出しているが，本提案手法では，PLSI 素性を用いることで用語の全体的な事例を考慮した用語抽出が可能になる．

表 1: 教師あり用語抽出結果

訓練辞書数	精度		再現率		F 値		最大 F 値		平均精度	
	線形	多項式	線形	多項式	線形	多項式	線形	多項式	線形	多項式
50	70.6	81.4	51.4	20.7	58.8	31.9	64.6	63.7	68.6	66.9
100	81.7	**92.1**	43.5	17.4	56.5	28.9	68.0	68.0	74.4	74.4
200	83.5	90.4	54.3	32.0	65.8	47.0	73.2	72.9	80.2	79.6
500	82.6	88.8	66.5	53.5	73.6	66.6	76.3	77.0	84.5	84.9
1000	82.6	87.7	70.3	63.3	75.9	73.5	78.3	79.3	86.3	87.4
3000	83.3	86.3	**78.0**	75.5	80.5	**80.6**	81.5	**82.3**	89.4	**90.6**

4 評価実験

4.1 評価に使用したテストデータ

提案手法を評価するテストデータとして，Web で公開されている日本電気株式会社の 2004 年の報道発表資料１年分（約 1.7MB）を用いた．そして，製品名・サービス名・機能名・技術名の計 3347 語の用語を抽出する．また，学習の補助として用いる Web から取得する文書数は，１つの用語候補あたり検索結果上位 10 件の Web 文書を取得し，その結果，約 685MB の文書集合を得た．

4.2 教師あり用語抽出実験

訓練辞書数とカーネル関数[3]を変化させながら，教師あり学習における本手法の精度を評価する．なお，訓練辞書は頻度に偏りがないように無作為に抽出し，各辞書数ごとに 10 通り作成し，辞書内の正例負例の比率はすべて 1:4 とした．そして，それらのテスト結果の 10 回平均をとることで，最終的な評価とした．また、訓練辞書に含まれる用語はテスト結果の評価対象としない．

4.3 半教師あり用語抽出実験

辞書数 200 のうち，教師あり学習実験結果の F 値に最も近い訓練辞書を用いて，半教師あり用語抽出を行う．ベースラインには，半教師あり SVM である Transductive Support Vector Machines(TSVM)[4] [9] と，素性を分割しない手法（手法４）を用意した。

手法１（提案手法） 用語内素性と用語外素性それぞれから分類器[5] を作成し，Co-Training を行う手法．用語外素性による分類器からは正例のみを５個抽出[6] し，用語内素性による分類器は，正例５個，負例 40 個を抽出する．各ラウンド毎の学習結果として全ての素性で学習した結果を出力する．この作業を 50 回繰り返す．[7]

手法２ 手法１と同じく素性を分割するが，初期文書集合のみを用いて PLSI 素性を作成した手法.

手法３ 手法１，２同じく素性を分割するが，PLSI 素性を加えない手法.

手法４ 素性を分割せずにブートストラップを行う手法. なお，作成する分類器が一つのため各ラウンド毎に，正例 10 個，負例 40 個を取得する．

4.3.1 評価方法

評価の算出方法は，用語の事例ごとに算出するのではなく，3.2 節で述べた評価関数から用語の代表スコアを算出し用語ごとに評価を行う．評価方法には精度と再現率，そして F 値を採用し，以下のように定義する．

$$\text{精度} = \frac{\text{システムが用語と判定したうちの正解用語数}}{\text{システムが用語と判定した数}} \times 100$$

$$\text{再現率} = \frac{\text{システムが用語と判定したうちの正解用語数}}{\text{テストデータ内に含まれる用語数}} \times 100$$

$$\text{F 値} = \frac{2 \cdot \text{精度} \cdot \text{再現率}}{\text{精度} + \text{再現率}}$$

また, 本研究では評価関数を用いて用語らしさのランキングを行うので, スコアの上位に用語が多数ある出力結果ほど, 高い評価が望まれる．そこで, スコア順にソートした場合の最高の F 値を評価基準として追加し，さらに情報検索の評価等で良く使われる平均精度を採用する．

4.4 実験結果

表１に教師あり学習の結果を示し，表２，図２，図３に半教師あり学習の結果を示す．図３は各ラウンドでの訓練辞書の精度をグラフ化したもので，訓練辞書の精度が高いほど頑健な学習ができていると言える．

教師あり学習においては訓練辞書数が少ない場合においてもまずまずの精度を得ることができ，辞書数を単純増加させるにつれて良い精度を得ることができた．半教師あり学習においては，提案手法が各ラウンドの訓練辞書追加の際に最もノイズを含まない学習が実現でき，F 値としても良い結果を示すことができた．この結果から，Web の事例を加え，さらに全ての事例の用語外素性を PLSI によって次元圧縮する本提案手法は成功しているといえる．

また，実験過程により得られた知見として，用語外素性のみでは技術名と製品名のような細かいクラスの差異を抽出することが大変難しいことがわかった．どちらの用語も，大変似た文脈で出現するパターンが多く，用語内素性でないと判別不能であるからだ．そのため，まずは，半教師あり学習で定義の大きなクラスの用語を抽出し，それらを抽出した後に教師あり学習

[3]線形カーネルと多項式カーネルを用いる．

[4]http://www.cs.cornell.edu/People/tj/svm_light/

[5]用語内素性と最終出力の分類器には線形カーネルを用い，用語外素性には多項式カーネルを用いた．

[6]正例の用例に対して，負例の用例は無限大であり，それらの中から最も負例である事例を抽出するのは大変困難である．また，用語外素性の分類器を作成する理由は，異なった視点を持つ分類器を作成する事が最大の理由である。そして，ノイズのある事例を訓練辞書に加えることが，Co-Training の実現を妨げることから負例の取得を除外した．

[7]訓練データを作成する文書集合は初期の文書集合のみを用いた．

表 2: 半教師あり用語抽出結果

手法	精度		再現率		F 値		最大 F 値		平均精度	
	初期	終了	初期	終了	初期	終了	初期	終了	初期	終了
手法１（PLSI:Web 文書 + 初期文書）	**78.4**	70.7	57.5	**79.1**	66.4	**74.7**	70.9	**74.7**	77.6	**82.3**
手法２（PLSI:初期文書）	**78.4**	69.1	57.5	77.6	66.3	73.1	70.8	73.3	77.6	80.7
手法３（PLSI なし）	**78.4**	69.0	57.5	78.6	66.3	73.5	70.8	73.6	77.6	80.4
手法４（素性の分割なし）	**78.4**	74.5	57.5	62.4	66.4	67.9	70.9	71.1	77.6	75.3
TSVM	**78.4**	64.5	57.5	55.2	66.4	59.5	70.9	60.3	77.6	63.0

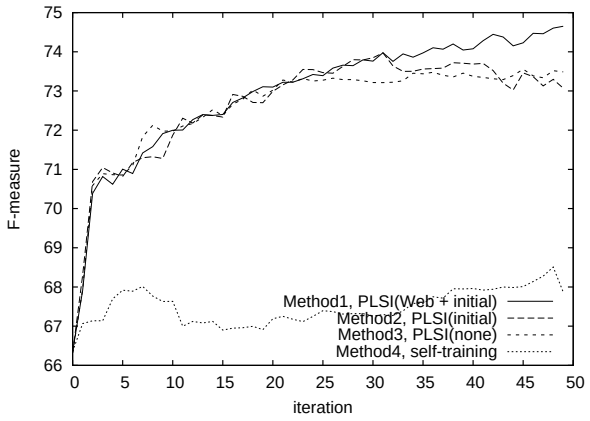

図 2: 各ラウンドにおける F 値の変化

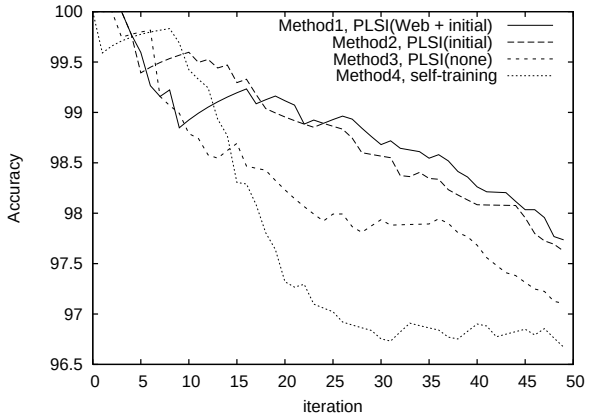

図 3: 各ラウンドにおける訓練辞書の Accuracy の変化

で細かいクラスの分類を行う手法が最も効率が良いと思われる．

5 おわりに

本稿では，Web サーチエンジンを利用することで用語の出現パターンを多数獲得し，さらに，用語の周辺部の統計的な分布から作成した確率モデルを素性とする半教師あり用語抽出手法を提案した．今回の実験においては，文書内に定型的な用語外素性のパターンが多かったことから，Web を用いない手法との明確な差が表れなかった．そこで，Web からさらに多くの用語事例を獲得することや，他の分野における用語抽出タスクに適用することで本提案手法の汎用性を実証すべきだと思われる．また今後は，Co-Training の各ラウンドに人手を介すことで頑健な学習を実現する能動学習手法を視野にいれつつ，より人手のかからない用語抽出手法を模索したい．

謝辞

本研究は日本電気株式会社インターネット研究所との産学共同研究の一環として行われました．研究を進めるにあたり，様々な御指摘，御助言をくださいました研究所の皆様に心から感謝致します．

参考文献

[1] A.Blum and T.Mitchell: Combining labeled and unlabeled data with co-training, *Proceedings of the Workshop on Computational Learning Theory (COLT'98)* (1998).

[2] D.Yarowski: Unsupervised word sense disambiguation rivaling supervised methods, *In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL'95)* (1995).

[3] F.Denis, A.Laurent, R.Gilleron and M.Tommasi: Text classification and co-training from positive and unlabeled examples, *In Proceedings of 20th International Conference on Machine Learning (ICML'03)* (2003).

[4] I.Muslea, S.Minton and C.Knoblock: Active + semi-supervised learning = robust multi-view learning, *In Proceedings of 19th International Conference on Machine Learning (ICML'02)* (2002).

[5] K.Nigam, A.McCallum, S.Thurn and T.Mitchell: Text classification from labeled and unlabeled documents using EM, *Journal of Machine Learning Research*, Vol. 39, pp. 103–134 (2000).

[6] K.Nigam and R.Ghani: Analyzing the effectiveness and applicability of co-training, *In Proceedings of of Ninth International Conference on Information and Knowledge (CIKM'00)* (2000).

[7] M.Collins and Y.Singer: Unsupervised models for named entity classification, *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP'99)* (1999).

[8] T.Hofmann: Probabilistic Latent Semantic Indexing, *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval (ACM'99)*, pp. 50–57 (1999).

[9] T.Joachims: Transductive Inference for Text Classification using Support Vector Machines, *In Proceedings 16th International Conference on Machine Learning (ICML'02)*, pp. 200–209 (1999).

[10] V.Vapnik: *Stastical Learning Theory*, Wiley-Interscience (1998).

[11] 宇津呂武仁, 颯々野学: ブートストラップによる人手低コスト日本語固有表現抽出, 情報処理学会研究報告, Vol. 2000, No. 86 (2000).

[12] 合原博, 宮田高志, 松本裕治: 医学生物学分野からの専門用語抽出・分類, 情報処理学会研究報告, Vol. 2000, No. 11 (2000).

[13] 国立国語研究所: 分類語彙表　-増補改訂版-, 大日本図書 (2004).