基本語ドメイン辞書と未知語ドメイン推定を用いたブログ自動分類

橋本 力* 黒橋 禎夫† 広川 仁‡ 横山 晶一*
山形大学大学院理工学研究科* 京都大学大学院情報学研究科† 山形大学工学部‡

1 はじめに

言葉の意味処理技術の深化に向けて、我々は基本語ドメイン辞書を構築した [1]。本研究では、その応用として、基本語ドメイン辞書を活用したブログ分類に取り組む。分類手法は、ブログ記事中の語にドメインを付与し、最も支配的なドメインにその記事を分類する、というシンプルなものである。しかし、本手法の特長は、辞書にある基本語だけでなく、未知語のドメインも動的に推定して分類の手掛かりとする点にある。また、機械学習を用いた手法とは異なり、大規模な文書集合を用意する必要がない。このような特長は、日々更新され、新語や造語が次々と現れるような、つまりブログ記事のような文書を分類する際に非常に有効である。本手法は、Yahoo! ブログ 600 記事を対象にした分類実験で正解率 94.0% (564/600) を達成した。また、分類の際に実行された未知語ドメイン推定の正解率は 77.2% (386/500) だった。

2 基本語ドメイン辞書

基本語ドメイン辞書では、JUMAN[3] の内容語約 30,000 語に対し、表 1 にある 12 ドメイン、あるいは＜ドメイン無し＞が付与されている。

表 1: ドメインとその手掛かり語の例

ドメイン	手掛かり語の例
文化・芸術	映画, 音楽, 文学, ···
レクリエーション	観光, 花火, 遊園地, ···
スポーツ	選手, 野球, 競技, ···
健康・医学	手術, 診断, 看護, ···
家庭・暮らし	育児, 家具, 住宅, ···
料理・食事	箸, 昼食, 喫茶, ···
交通	駅, 道路, 運転, ···
教育・学習	先生, 算数, 塾, ···
科学・技術	研究, 理論, 原子, ···
ビジネス	輸入, 市場, 経営, ···
メディア	放送, 記者, ＣＭ, ···
政治	司法, 税, 犯罪, ···

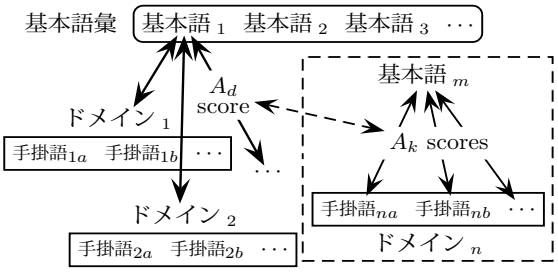

図 1: 各ドメインへの基本語の割り当て

2.1 構築手法

構築は §2.1.1、§2.1.2、§2.1.3、§2.1.4 の順に進む。

2.1.1 各ドメインへの手掛かり語付与

各ドメインに 20〜30 語ずつ、表 1 にあるような手掛かり語を人手で与える。

2.1.2 各ドメインへの基本語の割り当て

基本語と（＜ドメイン無し＞以外の）ドメインの間に関連度スコア（A_d スコア）を定義し、基本語を最も A_d スコアの高いドメインに割り当てる。A_d スコアは、基本語とドメインの各手掛かり語の間に定義される関連度スコア（A_k スコア）の上位 5 つを合計することで得られる。A_k スコアとして χ^2 に基づく指標を、コーパスとして Web を採用する [5]。共起頻度として、基本語と手掛かり語をクエリとした場合の検索エンジンヒット数を用いる。結局、基本語 w と手掛かり語 k の間の A_k スコアは以下のようになる。

$$A_k(w,k) = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

a = hits(w & k), b = hits(w)−a, c=hits(k)−a,
d = n−(a+b+c), n=日本語 Web ページ総数

$hits(q)$ は q をクエリとした場合のヒット数である。この段階で、各基本語は（＜ドメイン無し＞以外の）いずれかのドメインに割り当てられる。（図 1）

2.1.3 ＜ドメイン無し＞への再割り当て

割り当てられたドメインの A_d スコアが低い（≃ 設定された閾値以下の）基本語は＜ドメイン無し＞に再

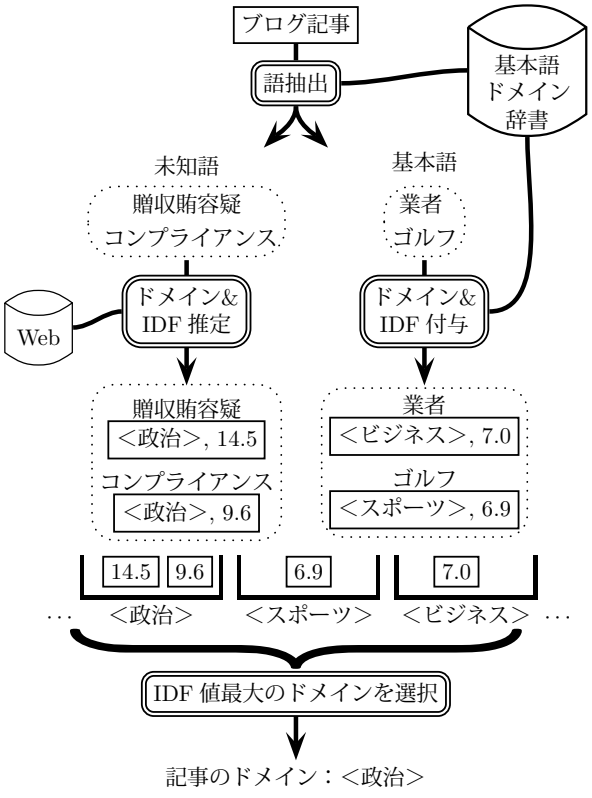

図 2: ブログ分類の全体像

割り当てされる。この段階で、適当と思われるドメインが全ての語に付与される。この段階でのドメイン付与正解率は **81.3% (309/380)** だった。

2.1.4 人手による修正

§2.1.3 までの結果を人手で修正し、完成させる。正解率が高いので作業負担はごく軽微である。

2.2 構築手法の特長

ドメイン選択の自由 この構築手法は表 1 の 12 ドメインとは独立である。つまり、好きなドメインを選んで独自のドメイン辞書を作ることができる。§2.1.1 の際、自分が選んだ各ドメインに適切な手掛かり語を付与すれば、あとの手順は全く同じである。

文書集合いらず 本手法は、多くの重要語抽出法とは異なり、文書集合を必要とせず、Web へのアクセスさえあればよい。この手軽さは、独自のドメイン辞書を作ろうとする際、非常に重要である。

3 ブログ自動分類

本研究ではブログ記事を表 1 のいずれかのドメインに分類する。分類は次のように進む（図 2）：❶ 記事中の語を抽出。❷ 各語にドメインと IDF 値を付与[1]。❸ ドメインごとに IDF 値を合計。❹ IDF 値合計が最も高いドメインを記事に割り当てる[2]。

IDF 値は次の定義に従う。

$$\text{語の IDF 値} = log\frac{\text{日本語 Web ページ総数}}{\text{語の Web ヒット数}}$$

語抽出❶として次の 3 通りを試みた：**1)** 基本語 **2)** 基本語と未知語 **3)** 基本語と未知語と複合名詞。**1)** と **2)** の場合は、複合名詞は分解され、その中の基本語が個別的に分類の手掛かりとして利用される。基本語のドメインと IDF は辞書から与えられるのに対し、未知語と複合名詞のドメインと IDF は、§4 で述べる手法により、動的に推定される。以後、未知語と複合名詞を合わせて「未知語」と呼び、区別する際は、前者を「単純未知語」、後者を「複合名詞未知語」と呼ぶ。

4 未知語ドメイン推定

未知語ドメイン推定は次の手順に従う（図 3）：❶ 未知語をクエリとして Web 検索を実行する。その際、検

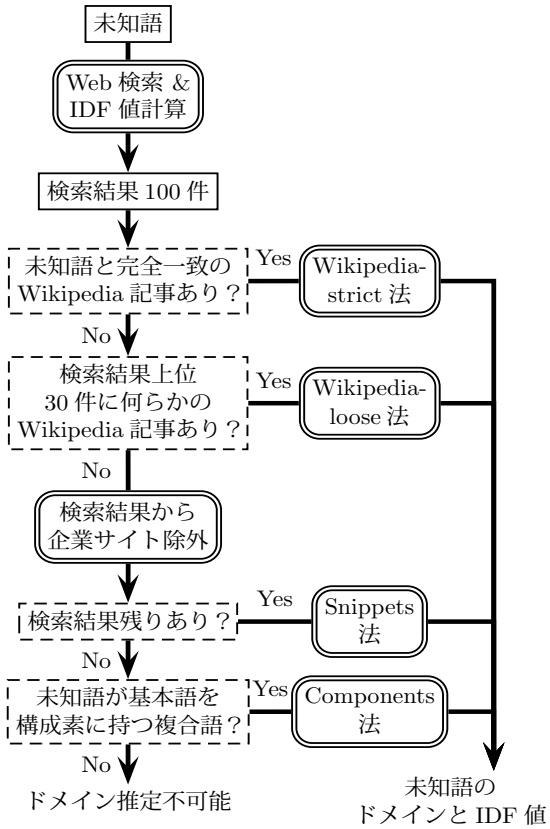

図 3: 未知語ドメイン推定手法の全体像

[1]実験に使用した辞書では、各語に対して、ドメインだけでなく IDF 値もあらかじめ付与しておいた。

[2]今回の実験では、IDF 値合計が最も高いドメインが＜ドメイン無し＞の場合は 2 番目のドメインを割り当てた。

索結果から得られるWebヒット数をもとに、その未知語のIDF値を計算する。❷ 検索結果上位100件の中に、その未知語とエントリが完全一致するWikipedia記事があれば、その記事を取得して、記事中の基本語を手掛かりにドメインを推定し、終了。(**Wikipedia-strict法**) ❸ もし未知語とエントリが完全一致するWikipedia記事が無ければ、検索結果上位30件の中から何らかのWikipedia記事を探し、もしあれば、その記事中の基本語を手掛かりにドメインを推定し、終了[3]。(**Wikipedia-loose法**) ❹ Wikipedia記事が全く見つからなければ、検索結果から企業の広告サイト等のスニペットを除外し、その残りのスニペットにある基本語を手掛かりにドメインを推定し、終了。(**Snippets法**) ❺ 企業サイトを全て削除すると何も残らない場合もある。その場合、未知語が基本語を構成素に持つ複合語なら、その構成語のドメインから未知語のドメインを推定して終了。(**Components法**) ❻ 検索結果にWikipedia記事も企業サイト以外のサイトも無く、また、未知語が基本語を構成素に持つ複合語でもない場合、失敗。

Wikipedia-strict法、Wikipedia-loose法、Snippets法、Components法について順に説明する。これらに共通するのは、手掛かりとなる記述(Wikipedia記事、検索結果のスニペット、複合語の構成語)にある基本語のドメインを調べ、最も支配的なドメインをその未知語のドメインとする、という発想である。

4.1 Wikipedia(-strict|-loose)法

Wikipedia-strict法とWikipedia-loose法の流れは次の通りである:① 検索結果をもとにWikipedia記事を取得。② 記事から基本語のみを抽出。③ 基本語ドメイン辞書を参照して、各基本語にドメインとIDF値を付与。④ IDF値合計が最も高いドメインを未知語に割り当てる。ただし、IDF値合計が最も高いドメインが<ドメイン無し>の場合、次の条件のもと、2番目にIDF値が高いドメインに割り当てる。

$$\frac{\text{2番目のドメインのIDF値}}{\text{<ドメイン無し>のIDF値}} > 0.15$$

ドメイン推定の所要時間は、Wikipedia-strict法、Wikipedia-loose法ともに約10秒である[4]。

[3]例えば未知語が「亀田兄弟」で、検索結果に「亀田兄弟」とエントリが完全一致するWikipedia記事がない場合、検索結果上位30件から何らかのWikipedia記事を探す。この例の場合、「亀田三兄弟」のエントリのWikipedia記事が見つかるので、その記事を取得し、ドメイン推定に利用する。

[4]使用した計算機はDell PowerEdge 830(Pentium Dプロセッサ3.00GHz、メモリ512MB)である。

4.2 Snippets法

Snippets法は、企業の広告サイト等のスニペットが削除された検索結果を入力として受け取る。企業の広告サイト等のスニペットは、未知語本来のドメインが何であれ、推定結果を<ビジネス>に偏らせてしまうため、削除する必要がある。本研究では図4にある語が2回以上現れたスニペットを企業サイトのものと判断する。Snippets法は、基本語抽出対象がスニペット

> 会社, 株式, 商品, 販売, 製品, 価格, 無料, 市場, 企業, ショップ, 通販, 事業, 発売, サービス, 法人, 店舗, 購入, 採用, 会員, 業務, 当社, 営業, 工業, ビジネス, 広告, 仕事, 出荷, 料金

図4: 企業サイトスニペットのキーワード

群である以外、Wikipedia法と同じである。

ドメイン推定の所要時間は6秒程度である[5]。

4.3 Components法

Components法は、基本語抽出対象が複合名詞未知語である以外、他と同じである。例えば「金融市場」なら「金融」と「市場」のドメインから全体を推定する。

ドメイン推定の所要時間は約4秒である[6]。

5 評価実験

5.1 評価データ

評価データとして、Yahoo!ブログ (`blogs.yahoo.co.jp`) からドメインあたり50記事、計600記事収集した。Yahoo!ブログでは、記事は投稿時にその著者によってYahoo!ブログカテゴリに分類される。実験に際して、ドメイン毎に適切なYahoo!ブログカテゴリを選び、そのカテゴリから50記事ずつ収集した[7]。

5.2 評価結果

5.2.1 ブログ分類の結果

正解率は表2の通りである。この結果は、手法が§3のような単純なものでも、基本語を対象に分類のための手掛かり(本研究ではドメイン情報)を整備することで、高い精度で分類が可能なことを示している。ま

[5]Wikipedia法の方が先に実行されるにも関わらず所要時間が長いのは、Snippets方が既に得られている検索結果を手掛かりとする一方、Wikipedia法では、新たにWebにアクセスし、Wikipedia記事を取得する手間がかかるためである。

[6]他の3つの手法よりも高速なのは、基本語抽出対象が他の手法よりもずっと小規模で、かつ、Wikipedia法のように新たにWebにアクセスする必要もないからである。

[7]不適切なカテゴリに分類されていたり、写真しかないような記事が約3割あった。それらは人手でより適切な記事と交換した。

表 2: ブログ分類正解率

上位 N	基本語	基本+単純未知語	基本+全未知語
1.	0.89	0.91	0.94
2.	0.96	0.97	0.98
3.	0.98	0.98	0.99
4.	0.99	0.99	1.00
5.	0.99	0.99	1.00

た、「基本 + 単純未知語」が「基本語」を上回り、「基本 + 全未知語」がその他 2 つを上回っていることが、未知語ドメイン推定がブログ分類に効果的であることを示している。

間違いの大半は、記事中の周辺的な話題を誤って取り上げてしまったことに起因する。例えば、観光旅行の記事では、その著者が旅行の交通手段に何度か言及したため、本来<レクリエーション>に分類されるべきところを、誤って<交通>に分類した。

5.2.2 未知語ドメイン推定の結果

正解率は 77.2% (386/500) だった。各手法の使用頻度と正解率は表 3 の通りである。最も精度の高

表 3: 未知語ドメイン推定手法の使用頻度と正解率

	使用頻度		正解率	
Wiki-str	0.146	(73/500)	0.85	(62/73)
Wiki-los	0.208	(104/500)	0.70	(73/104)
Snppts	0.614	(307/500)	0.76	(238/307)
Cmpnts	0.028	(14/500)	0.64	(9/14)
推定失敗	0.004	(2/500)		——

い Wikipedia-strict 法の頻度はそれほど高くないが、Wikipedia のエントリ数が増えるにつれ頻度が高くなり、結果、未知語ドメイン推定全体の正解率も高まることが期待できる。

推定成功例として、<ビジネス>として正しく推定された「デイトレ」(「デイトレード」の略) が挙げられる。複合名詞未知語の推定成功例には、「支持率」等が含まれる。その構成語である「支持」も「率」もそれ単体では<ドメイン無し>だが、全体としては<政治>だということが正しく推定された。

失敗例の多くは、<ドメイン無し>かそれ以外かの判断を誤ったものである。主なものとして市区町村名がある。本来<ドメイン無し>に属するが、ほとんどの地方自治体が行政等に関するホームページを開設しているため<政治>と誤判定される。

6 関連研究との比較

従来手法は機械学習を用いたものがほとんどである [4, 2]。一方本手法では、基本語のみを対象に分類の手掛かり (ドメイン) を整備しておくだけで済む。また、基本語ドメイン辞書を作るには、文書集合は必要なく Web へのアクセスさえあればよい (§2.2)。また、構築の際に要する手作業も軽微である (§2.1.4)。

本手法の分類体系は基本語ドメイン辞書のドメインに固定される。しかし、基本語ドメイン辞書のドメイン体系はユーザが目的に応じて自由に選べるため (§2.2)、必要なら目的に応じて辞書を作り直せばよい。また、文書分類が用いられる現場では、分類体系が頻繁に変更されるということは考えにくいため、いったん基本語ドメイン辞書を構築すれば済む。

もう一つの強みは未知語への対応能力である。本手法では、記事中に未知語を発見すると動的にそのドメインを推定する。その正解率も 77%と実用に耐えうる。この能力は、ブログのように、頻繁に更新され、新たな語が次々に現れるような文書の分類に非常に有効である。一方、従来手法で未知語に対応するには、訓練データを集め直す必要があり手間がかかる。しかもブログを対象とした場合、次々と未知語が現れるため、訓練データの更新を短い周期で行う必要がある。

7 おわりに

我々は基本語ドメイン辞書を活用して、機械学習 (そして文書集合) に頼らない、未知語にも柔軟に対応する文書分類手法を開発した。今後は、基本語ドメイン辞書と未知語ドメイン推定のさらなる応用として語義曖昧性解消に取り組む。

参考文献

[1] Chikara Hashimoto and Sadao Kurohashi. Construction of Domain Dictionary for Fundamental Vocabulary. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Poster*, pp. 137–140, Prague, Czech Republic, 2007.

[2] Robert E. Schapire and Yoram Singer. BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning*, Vol. 39, No. 2/3, pp. 135–168, 2000.

[3] 黒橋禎夫, 河原大輔. 日本語形態素解析システム JUMAN version 5.1 使用説明書. 京都大学大学院情報学研究科, 2005.

[4] 平博順, 春野雅彦. Support vector machine によるテキスト分類における属性選択. 情報処理学会論文誌, Vol. 41, No. 4, pp. 1113–1123, 2000.

[5] 佐々木靖弘, 佐藤理史, 宇津呂武仁. 関連用語収集問題とその解法. 自然言語処理, Vol. 13, No. 3, pp. 151–176, 2006.