基本語ドメイン情報の構築

橋本 力
山形大学工学部
ch@yz.yamagata-u.ac.jp

黒橋 禎夫
京都大学大学院情報学研究科
kuro@nlp.kuee.kyoto-u.ac.jp

1 はじめに

言葉の意味処理にとってシソーラスは不可欠の資源である。シソーラスは、単語間の上位下位関係という、いわば縦の関連を表現するものである。我々は意味処理技術の深化を目指し、縦の関連に加えて、単語が使用されるドメインという、いわば横の関連を提案する。例えば、単語が「教科書」「先生」ならドメインは＜教育・学習＞[1]であり、「庖丁」なら＜料理・食事＞、「メス」なら＜健康・医学＞である。ドメインを考慮することでより自然な単語分類が可能となる。例えば分類語彙表は、「教科書」は『文献・図書』、「先生」は『専門的・技術的職業』として区別するが、ドメイン上は両者とも＜教育・学習＞に属する。また、分類語彙表は「庖丁」も「メス」も『刃物』として同一視するが、両者はドメインにおいて区別される。

本研究では、基本語を対象に、ドメイン情報を半自動で構築した。本手法の強みは、語とドメインを対応づける際に、（多くの重要語抽出技術では不可欠な）ドメインごとの文書集合を必要としない点にある。また本研究では、基本語ドメイン情報を用いて、ブログの分類実験と未知語のドメイン推定実験を行った。

2 2つの問題

基本語ドメイン情報構築には 2 つの問題がある。1 つは世界を適切に分類するドメイン体系の設計であり、もう 1 つは文書集合を必要としないドメイン情報構築技術の開発である。

1 つ目の問題は、人間の外界認識の様式を明らかにするという難問である。本研究ではこの問題には深く立ち入らず、多くの人から合意が得られやすいと思われるシンプルなドメイン体系を採用した（表 1）。このドメイン体系は Open Directory Project (dmoz.org) 等の検索ディレクトリのカテゴリを参考にした。また、「人」や「青」のような特定のドメインに属さない語のために＜ドメイン無し＞も用意した。

[1]以後、本文中ではドメインを＜＞で囲んで表す。

表 1: 本研究のドメイン体系

文化・芸術	家庭・暮らし	科学・技術
レクリエーション	料理・食事	ビジネス
スポーツ	交通	メディア
健康・医学	教育・学習	政治

もう 1 つの問題は、あるドメインの文書集合からそのドメインのキーワードを抽出するといった重要語抽出技術が本研究には適用しにくいというものである。これは、表 1 のような、一般的・日常的な粒度のドメインの文書集合を集めることが困難なことに起因する。[2] 次節では、本研究で開発した、文書集合を必要としない基本語ドメイン情報構築手法について述べる。

3 基本語ドメイン情報構築手法

本手法のポイントは、基本語をドメインに割り当てるヒントとして、文書集合ではなく、少数の手掛かり語集合を用いる点にある。本手法の流れは次の通りである：①各ドメインへの手掛かり語付与（§3.1）②各ドメインへの基本語の割り当て（§3.2）③＜ドメイン無し＞への再割り当て（§3.3）④人手による修正（§3.6）。

3.1 各ドメインへの手掛かり語付与

表 1 のドメインに 20~ 30 語ずつ人手で手掛かり語を与える。手掛かり語は Web 高頻度語リストの上位から選ぶ。表 2 に手掛かり語の例を挙げる。

3.2 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

[2]当初我々は検索ディレクトリの登録ページを文書集合として利用した。しかし、登録ページの多くはいわゆる index ページで、統計的指標でキーワードを同定するには文章量が十分ではなかった。文章量を増やすため index ページのリンクを辿ってみたが、1 つのサイトから多くのページが収集されたため、ドメインのキーワードというより、サイトのキーワードというべき語が抽出された。他に、関連性の薄い広告リンクを辿ってしまうという問題もある。

表 2: 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, ···

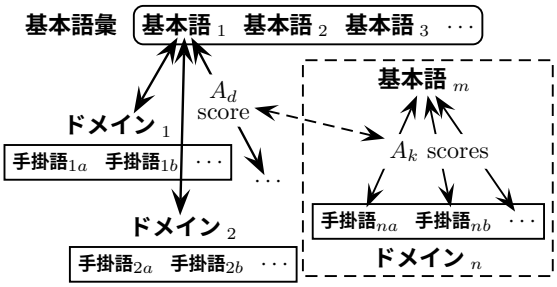

図 1: 各ドメインへの基本語の割り当て

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。本研究では [6] に従い、コーパスにおいてよく共起する語ほど関連度が高いという前提のもと、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。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。

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

ただし n は日本語 Web ページ総数で[3]、a から d はそれぞれ以下のようになる。

$$a = hits(w \ \& \ k) \quad b = hits(w) - a$$
$$c = hits(k) - a \quad d = n - (a+b+c)$$

$hits(q)$ 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 参照）

[3]我々は 10,000,000,000 を n とした。

3.3 ＜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 スコアが低い基本語は＜ドメイン無し＞に再割り当てされる。ここで A_d スコアが低いかどうかを決める閾値が必要となる。我々が行った予備調査によると、検索エンジンヒット数が多い基本語ほど閾値を高めに設定する必要がある。

そこで、ヒット数に応じた適切な閾値を与える関数を次の手順で得た：①< 基本語, ヒット数, 割り当てられたドメインの A_d> の 3 つ組をヒット数の降順に並べる[4]。② 3 つ組の集合を 130 のヒット数セグメントに分割する。③各セグメントから、＜ドメイン無し＞に属すべき基本語を含む 3 つ組とそれ以外の 3 つ組をそれぞれ 5 つ手作業で抽出する[5]。④セグメントごとに、＜ドメイン無し＞に属すべき基本語を含む 3 つ組とそれ以外の 3 つ組を分離する A_d スコアの値を同定する。この値が当該ヒット数セグメントにおける閾値となる[6]。⑤ヒット数と閾値の関係を最小二乗法により 1 次関数で近似する。この 1 次関数がヒット数に応じた適切な閾値を与える関数である。（図 2 参照）

3.4 ドメイン割り当ての性能評価

§3.1 から §3.3 で述べた手法のドメイン割り当て正解率を測定した。その際、基本語として Juman [5] にある内容語の名詞と動詞、計 26,658 語を使用した。ドメイン割り当て結果から基本語–ドメインのペアを 380 組抽出し、そのうち何%が正解か調べた。比較のためベースラインも用意した。ベースラインは、全ての基本語を＜ドメイン無し＞とした場合の正解率である。これは、予備調査の段階で、基本語の半分以上が＜ドメイン無し＞と判定されることがわかったためである。

結果として、81.3% (309/380) の正解率を得た。一方ベースラインの正解率は 69.5% (264/380) だった。

3.5 複数ドメインの割り当て

ある基本語は複数のドメインに属す。例えば「大学院」なら＜教育・学習＞と＜科学・技術＞の両方に属すものと考えられる。しかし、上述の手法は 1 語を 1 つのドメインにしか割り当てないように設計されている。本節では、1 語を複数のドメインに割り当てることが可能な、上述の手法の拡張版について述べる。

語を、A_d スコアが最も高いドメインだけでなく、以下の 2 つの条件を満たすドメイン全てに割り当てる：①そのドメインの A_d スコアが §3.3 で述べた閾値以上

[4]§3.2 の段階でこれら 3 つ組が全て得られていることに注意。
[5]通常、前者より後者の 3 つ組の方が A_d スコアが高い。
[6]この段階で、（セグメントによって表された）ヒット数とその閾値のペアが 130 組得られる。

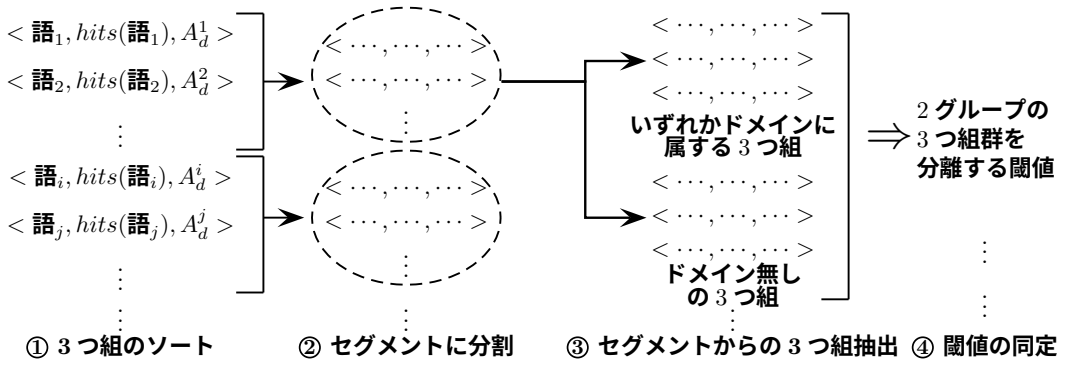

図 2: ＜ドメイン無し＞への再割り当て：1 から 4 まで

表 3: 基本語ドメイン情報の語の内訳

ドメイン	%	ドメイン	%
文化・芸術	2.4	教育・学習	1.6
レクリエーション	0.8	科学・技術	1.1
スポーツ	1.6	ビジネス	4.2
健康・医学	2.9	メディア	0.6
家庭・暮らし	3.2	政治	4.5
料理・食事	3.3	ドメイン無し	72.7
交通	1.1		

表 4: B_{random} の内訳

ドメイン	#	ドメイン	#
文化・芸術	4	料理・食事	4
レクリエーション	1	ビジネス	12
スポーツ	3	ドメイン無し	5
健康・医学	1		

である。②そのドメインの A_d スコアが最も高い A_d スコアに十分近い。②は次のように定式化される。

$$\frac{\text{最も高い } A_d - \text{そのドメインの } A_d}{\text{最も高い } A_d} < 0.01$$

この手法により、基本語–ドメインの組が 814 組増えたが、正解率は 78.6% (308/392) に落ちた[7]。

3.6 人手による修正

人手修正の際の指針の 1 つとして、複数ドメインに割り当てるべき基本語を、それらのドメインに同程度に関連するものに限定するというものを設けた。これには、基本語ドメイン情報をなるべくシンプルなものにするという狙いがある。表 3 は完成した基本語ドメイン情報の語の内訳である。

4 ブログ分類実験

評価の一環として、基本語ドメイン情報を用いてブログ記事をドメインごとに分類する実験を行った。

4.1 分類手法

①ブログ記事から基本語を抽出。②抽出された基本語をドメイン情報によりドメインに分類。③分類された基本語の数の降順にドメインをソート。④記事を最上位のドメインに分類。最上位が＜ドメイン無し＞の場合[8]は、次の条件のもと、2 番目のドメインに分類。

$$\frac{\text{2 番目のドメインの基本語数}}{\text{＜ドメイン無し＞の基本語数}} > 0.03$$

4.2 データ

ブログ記事群として $B_{controlled}$ と B_{random} の 2 つを用意した。$B_{controlled}$ として、1 ドメイン 3 記事、計 39 記事を次の手順で収集した：①Google Blog Search (www.google.co.jp/blogsearch) にドメインの手掛かり語を 1 語クエリとして与える[9]。②検索結果上位から次の 2 条件を満たす 3 記事を収集：2-1. 十分な文章量がある。2-2. 当該ドメインに属すと人間が明確に判断できる。B_{random} として、30 記事を Web からランダムに集めた。表 4 はその内訳である。

両記事群ともに、分類前に、記事中の周辺的なコンテンツ（プロフィールや広告等）を人手で削除した。

4.3 実験結果

結果、$B_{controlled}$ は 89.7% (35/39) の正解率で、B_{random} は 76.6% (23/30) の正解率で分類できた。

5 未知語のドメイン推定

基本語ドメイン情報を用いて未知語のドメインを推定する方法について述べる。

[7] §3.6 の人手による修正では、その正解率の高さから、複数ドメイン版ではなく、単数ドメイン版の手法を使用した。複数ドメインに割り当てるべき基本語には人手による修正の段階で対応した。

[8] 平均で、記事中の語の 69%が＜ドメイン無し＞に分類されるので、多くの記事がこの場合に当てはまる。

[9] ＜ドメイン無し＞の場合、「日記」をクエリとした。

表 5: 未知語ドメイン推定の正解数（10 語中）

ドメイン	#	ドメイン	#
文化・芸術	7	交通	7
レクリエーション	4	教育・学習	9
スポーツ	9	科学・技術	6
健康・医学	9	ビジネス	9
家庭・暮らし	3	メディア	2
料理・食事	7	政治	9

5.1 推定方法

①未知語をクエリとして Web を検索[10]。②検索結果上位 30 文書を収集。③各文書を §4.1 の手法でドメインに分類。④分類された文書数の降順にドメインをソート。⑤未知語を最上位のドメインに割り当てる。

5.2 実験手順

①各ドメイン（＜ドメイン無し＞以外）につき 10 語ずつ、基本語ドメイン情報からランダムに選定。②選ばれた各語に対し、その語を基本語ドメイン情報から除去した上で、§5.1 の手法でその語のドメインを推定。

5.3 実験結果

表 5 に、10 語中、ドメインが正しく推定された未知語の数を挙げる。全体の正解率は 67.5% (81/120) である。表 5 を見ると、＜レクリエーション＞、＜家庭・暮らし＞、＜メディア＞の結果の悪さが目立つ。＜メディア＞の場合、それに属す語のドメインに関する曖昧性が結果の悪さの主な要因だった。例えば、「中継」は＜メディア＞に属すが、その語はスポーツ等の文脈においてもよく使用される。＜レクリエーション＞と＜家庭・暮らし＞の結果の悪さは、Web を活用するという本手法の特徴が裏目に出たものと言える。つまり、それらのドメインに属する語、例えば「観光」や「シャンプー」等は、レクリエーションや家庭・暮らし関係のサービスや商品を扱っている企業（ビジネス）の Web サイトで頻繁に現れるため、誤って＜ビジネス＞と判定されてしまうことがある。

6 関連研究

既存のドメイン資源として HowNet [2] と WordNet(2.0) [3] があるが、（広く利用可能な）日本語用のドメイン資源は無かった[11]。ドメイン情報構築手法に関しても、従来は LDOCE や WordNet の情報を利用する手法 [4, 1] がほとんどで、そのような資源が無い言語には適用できない。また、重要語抽出技術も、本研究で対象としている一般的・日常的ドメインの文書集合を得るのが困難なため、適用が難しい。つまり本研究の貢献は、日本語用のドメイン資源を構築したことと、WordNet のような資源も文書集合も必要ないドメイン情報構築手法を開発したことの 2 点である。

[10] 本研究では Yahoo! JAPAN を使用した。

[11] 辞書の語義文に一部ドメイン情報に相当する記述があるが、本研究で対象としている一般的・日常的ドメインはカバーしていない。

7 まとめ

本研究では、[6] に基づき、基本語ドメイン情報の半自動の構築法を提案し、26,658 語の名詞と動詞を対象として、実際にドメイン情報を構築した。本手法には WordNet のような資源も文書集合も必要ない。基本語ドメイン情報を用いたブログ分類実験では、両データセットにおいて良好な結果を得た。基本語ドメイン情報を用いた未知語ドメイン推定実験では、ドメインに関して曖昧な語や、企業の Web サイトで頻出する語を除いては、良好な結果を得た。

日常的に使われる複合語のドメインの扱いは今後の課題である。例えば「源泉」は＜ドメイン無し＞だが、「源泉徴収」となると＜政治＞として扱うべきである。

参考文献

[1] Eneko Agirre, Olatz Ansa, and David Martinez. Enriching wordnet concepts with topic signatures. In *Proceedings of the SIGLEX Workshop on "WordNet and Other Lexical Resources: Applications, Extensions, and Customizations" in conjunction with NAACL*, 2001.

[2] Zhendong Dong and Qiang Dong. *HowNet And the Computation of Meaning*. World Scientific Pub Co Inc, 2006.

[3] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[4] Joe A. Guthrie, Louise Guthrie, Yorick Wilks, and Homa Aidinejad. Subject-Dependent Co-Occurence and Word Sense Disambiguation. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 146–152, 1991.

[5] 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.

[6] 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.