基本語ドメイン辞書の構築と未知語ドメインの推定を用いたブログ自動分類手法への応用

橋本　力　　黒橋禎夫
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

発表：林華

1　はじめに

- 言葉の意味処理にはシソーラスが不可欠
- シソーラスは単語間の上下関係で、縦関連

  横の関連がない
- 単語のドメイン、いわば横の関連を提案
- ドメイン辞書構築手法は半自動プロセス
  - 全自動は技術的に困難
  - 人手はコスト的、一貫性、保守性望ましくない

2　二つの問題

- 世界を適切に分類するドメイン体系の設計
  - 人間の外界認識の様式を明らかにする難問
  - Open Directory Project(http://www.dmoz.org)
  - <ドメイン無し>も用意した
- 文書集合なしでの基本語ドメイン辞書構築
  - 一般的・日常的な粒度のドメインの文書収集困難
  - 文章量が不十分
  - ドメインのキーワードの誤抽出の問題

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

３　基本語ドメイン辞書構築手法

- 各ドメインへの手掛かり語の付与
- 各基本語へのドメインの割り当て
- <ドメイン無し>の割り当て
- 人手による修正

注意：以下で述べる構築手法は特定のドメイン体系に依存しない。

3.1 各ドメインへの手掛かり語の付与

- 手掛かり語はWeb高頻度語リスト上位選ぶ
- 判断に迷う語は無視し、明確な語だけ選ぶ
- 表1のドメインに20~30語ずつ人手で付与

注意：本研究と異なるドメイン体系を採用した場合は、異なる手掛かり語を独自に収集する必要がある。しかし、以下に述べるその後のプロセスが全く同じである。

表2: 手掛かり語の例

文化・芸術	写真、映画、音楽、アニメ、曲、デザイン、文化、展示、映像、美術、芸術、…
レクリエーション	遊園地、ゲーム、遊ぶ、旅行、温泉、観光、旅、趣味、パーティー、おもちゃ、…
スポーツ	選手、試合、スポーツ、野球、サッカー、レース、ボール、スキー、対戦、…
…	…

3.2 各基本語へのドメインの割り当て

- 基本語とドメインの間の関連度スコア（A_d）を定義
- 基本語とドメインの各手掛かり語関連度（A_k）計算
- 関連度上位の5つの合計をA_dとする
- A_kスコアとして$\chi 2$を、コーパスはWebを採用
- コーパスにおいてよく共起する語ほど関連度が高い

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

$$a = hits(w \ and \ k) \qquad b = hits(w) - a$$

$$c = hits(k) - a \qquad d = n - (a+b+c)$$

3.3 <ドメイン無し>の割り当て

- <基本語,ヒット数,A_d>をヒット数の降順並べる
- 3つ組の集合を130のヒット数セグメントに分割
- 各セグメントから、<ドメイン無し>とそれ以外のドメインに属すべき単語をそれぞれ5つ抽出
- セグメントごとに、<ドメイン無し>に属すべき組とそれ以外組を分離するA_dスコアを同定し、閾値を決定
- 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

3.4 ドメイン割り当ての性能評価

- ドメイン割り当てから基本語ードメインのペアを380組抽出し、正解率を調べた。
- 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。
- 結果として、81.3%（309/380）正解率を得て、ベースラインの正解率は69.5%（264/380）だった

3.5 複数のドメインの割り当て

- 複数にドメインに割り当てが可能
  - 例：大学院　<教育・学習>、<科学・技術>

    円高　　<ビジネス>、<政治>
- 語を次の条件満たすドメイン全てに割り当て
  1. そのドメインのA_dが3.3で述べた閾値以上
  2. そのドメインのA_dが最も高いA_dに十分近い

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

3.7 人手による修正

- 複数ドメイン版ではなく、単一ドメイン版使用
- 複数のドメインに属する語の扱い：それら複数のドメインに同程度に関連するもののみに限定
  - 例：大学院　<教育・学習>、<科学・技術>
    登山　<レクリエーション>、<スポーツ>
    微分　<教育・学習>で、<科学・技術>ではない
- 多義語の扱い：各語義に対応するドメインを割り当て
  - 例：ボール　<スポーツ>、<料理・食事>
- <ドメイン無し>の判定基準：意見分かれそう語や5つ以上のドメインに属する語
  - 例：委員、組織など

4　基本語ドメイン辞書の詳細

ドメイン	%
文化・芸能	4.7
レクリエーション	1.0
スポーツ	2.5
健康・医学	3.4
家庭・暮らし	5.4
料理・食事	3.9
交通	1.6

ドメイン	%
教育・学習	2.3
科学・技術	1.4
ビジネス	3.6
メディア	0.7
政治	6.2
ドメイン無し	**63.4**

- 完成した基本語ドメイン辞書をJUMANに組み込む
  - 例：研究室のゼミで先生と議論した。

    研究　けんきゅう　研究　名詞　6　サ変名詞2*0*0　ドメイン：科学・技術

    室　しつ　室　名詞　6　普通名詞　1*0*0

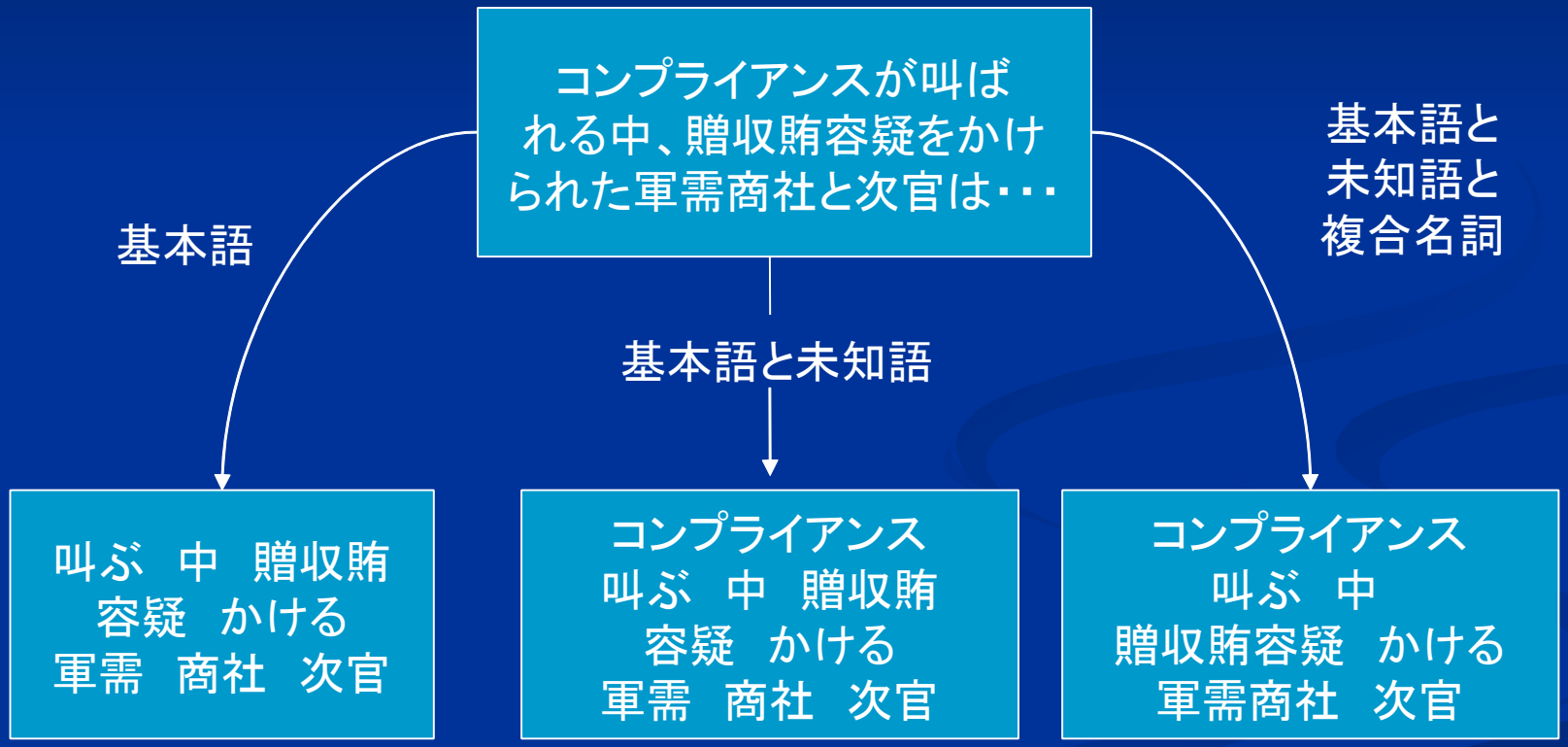
    ・・・

５ ブログ自動分類への応用

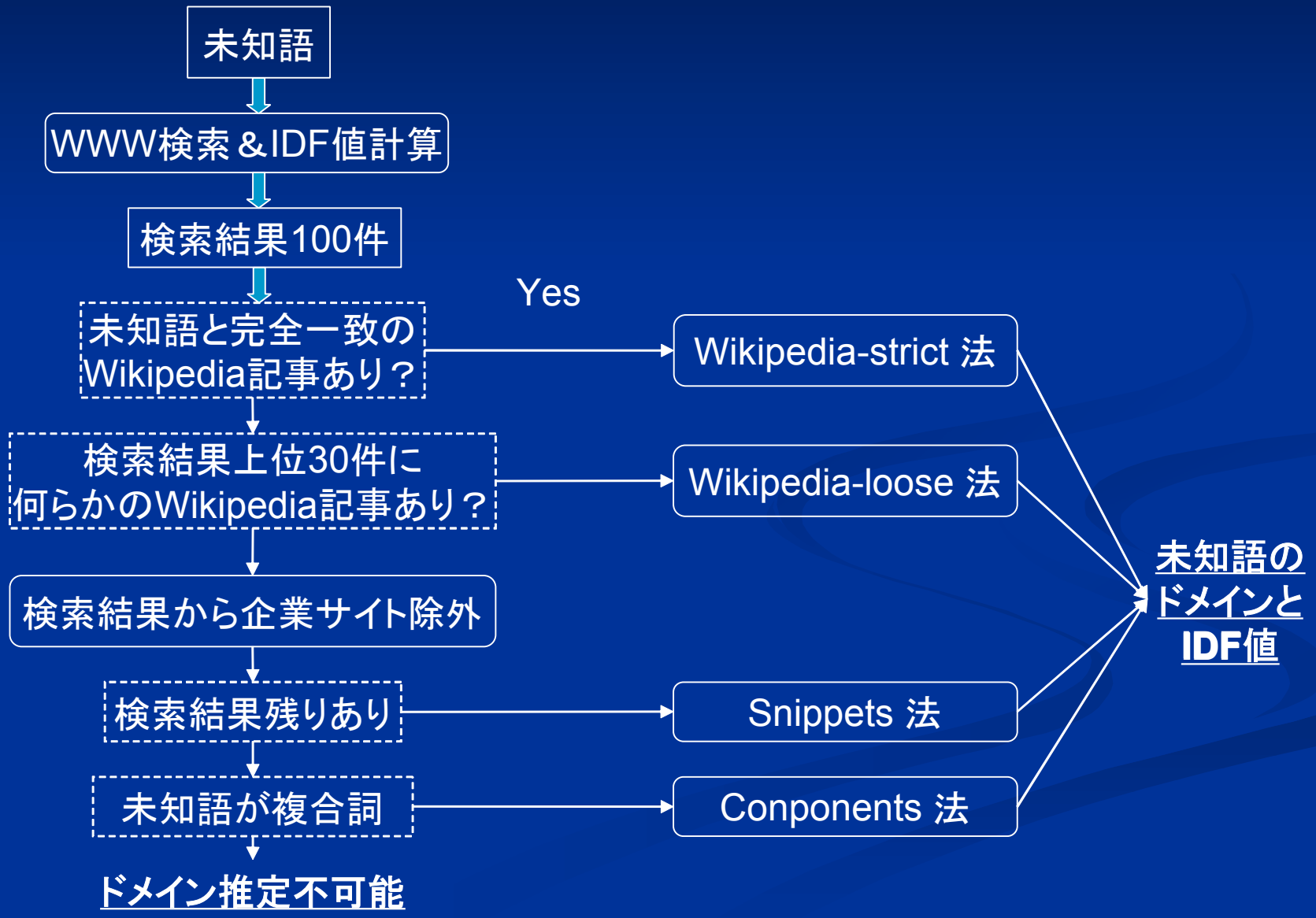
1. 記事中の語を抽出
2. 各語にドメインとIDF値を付与
3. ドメインごとに、割り当てられたごのIDF値を合計
4. IDF値合計の最も高いドメインを記事に割り当てる
5. IDF値合計の最も高いドメインが<ドメイン無し>の場合は2番目のドメインに割り当てる

$$\mathrm{IDF} = \log \frac{\mathrm{Web}}{\mathrm{Web}}$$

未知語について


コンプライアンスが叫ばれる中、贈収賄容疑をかけられた軍需商社と次官は・・・
基本語
基本語と未知語
基本語と
未知語と
複合名詞
叫ぶ　中　贈収賄
容疑　かける
軍需　商社　次官
コンプライアンス
叫ぶ　中　贈収賄
容疑　かける
軍需　商社　次官
コンプライアンス
叫ぶ　中
贈収賄容疑　かける
軍需商社　次官

6　未知語ドメインの推定

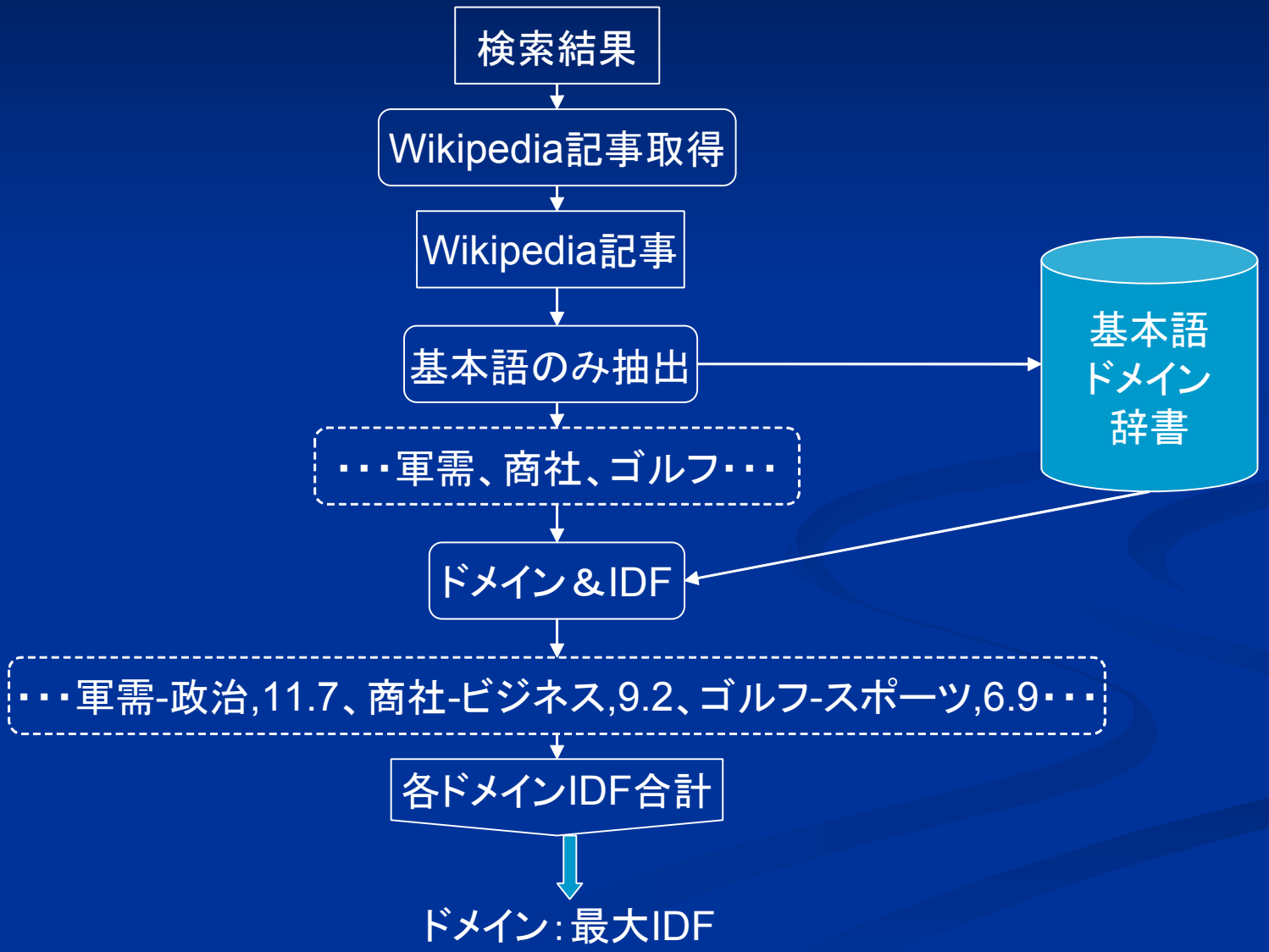


6.1 Wikipedia(-strict|-loose)法

1. 検索結果を基に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、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か次の条件を満たす2番目ドメインを未知語に割り当てる。

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

Wikipedia手法による未知語ドメイン推定



6.2 Snippets 法

1. 企業の広告サイトやスニペットなどを除去
   - 会社、株式、商品、販売、製品、価格、無料、市場、通販、事業、・・・
2. 検索結果から基本語のみを抽出
3. 基本語ドメイン辞書を参照し、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か次の条件を満たす2番目ドメインを未知語に割り当てる。

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

6.3 Components 法

1. 未知語を構成する基本語を抽出
2. 基本語ドメイン辞書を参照し、各基本語にドメインとIDF値を付与
3. IDF値合計が最も高いドメインか次の条件を満たす2番目ドメインを未知語に割り当てる。

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

7 ブログ分類と未知語ドメイン推定の評価実験
7.1.1 実験条件—評価データ

- Yahoo!ブログ(http://blogs.yahoo.co.jp)から、各ドメインにつき50記事ずつ、600記事
- 記事が投稿時に著者によりカテゴリに分類

ドメイン	Yahoo!ブログカテゴリ
文化・芸術	エンターテインメント＞映画 エンターテインメント＞音楽 エンターテインメント＞芸能人、タレント 芸術と人文＞芸術、アート 芸術と人文＞舞台、演劇 生活と文化＞祝日、記念日、年中行事
レクリエーション	エンターテインメント＞テーマパーク 趣味とスポーツ＞レジャー 趣味とスポーツ＞趣味
スポーツ	趣味とスポーツ＞スポーツ
健康・医学	健康と医学
…	…

7.1.2 評価方法

- ブログ分類手法
  - 利用する語
    1. 基本語のみ
    2. 基本語＋単純未知語
    3. 基本語＋単純未知語＋複合名詞
  - IDF値合計がトップだけではなく、上位5位まで
  - IDF値ではなく、ドメインごとの語数（利用する語は3だけ）
- 未知語ドメインの推定手法
  - 評価データ中の未知語約12000語の推定結果からの500件について正解率を調べ
  - 複数ドメイン場合、一つが推定されれば正解
  - 未知語ドメイン推定で使用した各手法の使用頻度と正解率を測定

7.2 ブログ分類結果

ブログ分類正解率（手作業修正あり辞書）

上位N位	基本語	基本語＋単純未知語	基本語＋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

ブログ分類正解率（手作業修正無し辞書）

上位N位	基本語＋全未知語
1	0.82
2	0.90
3	0.91
4	0.92
5	0.92

ブログ分類正解率（語数を利用）

上位N位	基本語＋全未知語
1	0.91
2	0.97
3	0.99
4	1.00
5	1.00

7.3 未知語ドメイン推定結果

各手法	使用頻度	正解率
Wikipedia-strict	0.146 (73/500)	0.85 (62/73)
Wikipedia-loose	0.208 (104/500)	0.70 (73/104)
Snippets	0.614 (307/500)	0.78 (239/307)
Components	0.028 (14/500)	0.64 (9/14)
推定不可能	0.004 (2/500)	
総合	0.996 (498/500)	**0.766** (383/500)