テキストの分野に応じた意味を表示する辞典選択システムの評価

關 博之　　岩下 志乃
東京工科大学コンピュータサイエンス学部
iwashita@cs.teu.ac.jp

1　はじめに

インターネットの普及やWWW上のツールの種類・品質の向上により，言葉や事柄について調べる際に，Web上のツールを用いて調べる機会が増えている．検索エンジンによって得られるWeb情報は，リアルタイムで量が豊富だが，質が低い．その一方，人手で編集される辞典や百科事典では，統制のとれた質の高い情報を利用できるが，情報量の制限や著者の視点に偏る．

既存研究において，情報の質の向上として要約や未知語抽出，テキスト分類・関係等についての手法が提案されている [1][2]．また，インターネット上では情報量が限られる専門的用語に対し，特許情報から用語の意味情報を抽出するシステムの実現がされてきた [3]．しかし，要約における情報の欠落の可能性や新語，複合名詞における固有名詞・普通名詞の区別，用語分類の生成の際に利用する情報源に依存する等の問題が挙げられている．

本研究では，テキスト内で使用されている用語に対して，そのテキストの分野に応じた適切な意味を複数のWeb辞典から選択して表示するシステムの構築を目的とする．

表1に示すように，利用されるテキストの分野の違いにより意味が変化する用語が存在する．このような場合に，1つの辞典では間違った意味を表示してしまう可能性がある．そこで本システムでは，テキストに含まれる名詞と複合名詞と，予め分野ごとに分けられたWeb辞典とのマッチングを行う．分野ごとに集計した計算結果に基づき，最も点数の高い分野をテキストの分野であるとし，辞典の長所である「統制の取れた質の高い情報」と，Web情報の長所である「リアルタイムで情報量が豊富」を合わせたWeb辞典の情報を利用し，テキストに対し用語の適切な意味を表示する．評価として，テキストの分野判定の精度と表示された意味の妥当性を調べ，システムの有用性について議論する．

表 1: テキストの分野における用語の意味の変化

用語	テキストの分野		
	一般的用法	IT	芸人
継承	身分・仕事財産などを受け継ぐ	他のクラスで定義されている動作を再利用…	特になし
フルーツポンチ	デザートの一種	特になし	吉本興業東京本社のお笑いコンビ

2　Web辞典について

2.1　Web辞典の信頼性

存在するWeb辞典の多くは安易に情報の編集ができる形式ではない．加えて，特定の専門的知識のある企業や団体がサービスの一環として提供しているWeb辞典もあることから，信頼性の高い情報であるといえる．

フリー百科事典であるWikipediaのように，閲覧者が説明文を自由に書き換えられるサービスの場合，説明文に対する信頼性に確証は得られない．しかし，自由に用語に対する記述行えるという観点を考慮すると，リアルタイムに限定的な用語についても意味取得を行える長所を持つ．この事を評価し，本システムでは自由書き込み可能なWeb辞典である場合でも利用している．

2.2　Weblioについて

本システムでは，多くのWeb辞典を利用すると共に利用するWeb辞典を特定のカテゴリに分けておくことが望ましい．これを踏まえ，複数の専門辞書や国語辞典，百科事典を横断的に検索することができる

統合型のオンライン辞書である「Weblio」を用いる．Weblio では，多種にわたる Web 辞典の情報を統一的な表示形式で利用，および統一的な HTML 形式で解析することが可能であり，手軽に多くの情報を利用できるメリットがある．2012 年 1 月 8 日時点では 613 種の辞書や事典を検索することができる．

また，Weblio では 19 種のカテゴリに Web 辞典を振り分けている．本システムでは，この 19 種のカテゴリとそれに属する特定の Web 辞典を用いる．特定の Web 辞典とは，1)Weblio 辞書，2)Weblio 類語辞典，3)Weblio 英和和英辞典，4)Weblio 日中中日辞典，5)Weblio 手話辞典，6)Weblio モバイルなどのサービスの中から，国語辞典や専門用語辞典を中心とする「Weblio 辞書」で利用できる一部 Web 辞典の事である．利用するカテゴリ類について，表 2 に示す．

表 2: カテゴリ類と利用する Web 辞典数

カテゴリ名	所属 Web 辞典数
ビジネス	40
業界用語	24
コンピュータ	28
電車	40
自動車・バイク	41
船	26
工学	40
建築・不動産	23
学問	35
文化	42
生活	19
ヘルスケア	32
趣味	25
スポーツ	14
生物	33
食品	20
人名	23
方言	23
辞書・百科事典	13
Web 辞典の総数	541

3　システム概要

3.1　処理の流れ

図 1 に本システムの概要を示す．

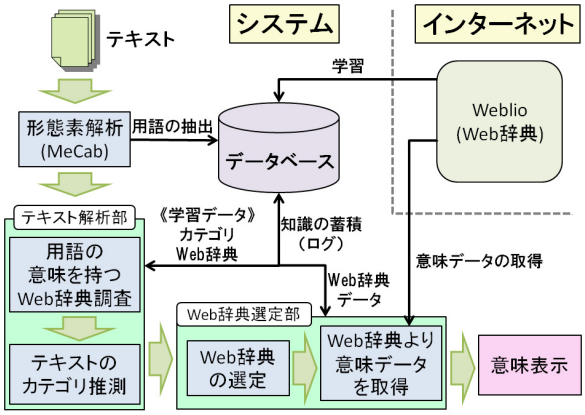

図 1: システム概要

入力されたテキストから，日本語形態素解析エンジンである MeCab を用いて形態素情報を取得する．取得した形態素情報から名詞と複合名詞を取得し，加えて取得した名詞の意味を持つ Web 辞典を調べる．求められた出現名詞の情報より，テキストのカテゴリ推測を行う．テキストの分野については，Weblio で用いられている 19 種のカテゴリに割り当て利用する．カテゴリ推測の結果に基づき，検索対象の名詞についてシステム内で利用する Web 辞典を選定し，適切な意味を持つ可能性が一番高い Web 辞典の意味情報を優先的にユーザに出力する．

次節で，テキスト解析部と Web 辞典選定部について説明する．

3.2　テキスト解析部

対象テキストより取得された名詞と複合名詞に対し，名詞の意味を持つ Web 辞典を Weblio より割り当てる．その際，Web 辞典が属するカテゴリ情報も共に取得し，テキストの特徴量（以下，ポイント）として次節で扱う．

次に，名詞と複合名詞の Web 辞典の情報を用いて，対象テキストのカテゴリを推測する．本システムでのカテゴリ推測には，式 (1) を用いて，テキスト中での名詞の出現回数と，名詞の意味を持つ特定のカテゴリに属する Web 辞典の数の積を各カテゴリごとに集計し，得られた数値よりテキストのカテゴリを定める．テキストの分野については，Weblio で用いている 19 種のカテゴリに当てはめて推測する．

$$T(C) = \sum_{n=1}^{M} (Count(Nn) \cdot NCD(Nn, C)) \qquad (1)$$

なお，$T(C)$ はテキスト T のカテゴリ C のポイント，Nn は処理に用いられる名詞，M はテキストの全出現名詞数，C はカテゴリ，$Count(Nn)$ は名詞 N の出現総数，$NCD(Nn, C)$ は名詞 Nn の意味を持つカテゴリ C に属する Web 辞典の総数である．

本システムでは，カテゴリ「辞書・百科事典」に属する汎用的な Web 辞典においては，優先度を 2 位または 3 位に再配置し，専門性の強い Web 辞典を優先的に利用するように考慮した．例として，コンピュータ関連のテキストについてカテゴリ推測した結果を表 3 に示す．

表 3: カテゴリ推測の結果

順位	カテゴリ名	ポイント
1	コンピュータ	159
2	学問	130
3	辞書・百科事典	750
4	建築・不動産	33
5	業界用語	29
6	ビジネス	21

3.3 Web 辞典選定部

検索する用語の意味表示において，利用する Web 辞典の選定をする．用語として取得された名詞と複合名詞の意味を持つ全 Web 辞典に対し，テキストのカテゴリ推測の結果と各 Web 辞典が属するカテゴリ情報を用いて，テキストに適した意味を持つ可能性の高い順に優先順位を定める．

Web 辞典の選定結果より，優先度が 1 位と定められた Web 辞典から用語の意味を取得し，ユーザに出力する．また 1 位でない Web 辞典に対しても，意味取得を行うことは可能である．

4 評価

4.1 カテゴリ推測の精度評価

本システムでは，テキスト中の名詞情報に基づきカテゴリ推測を行うことから，専門用語を多用しているテキストにおいてその特徴が表れると考えられる．これを踏まえ，専門性が高く専門用語を多用していると思われるカテゴリ「株（ビジネス）」「コンピュータ」「ヘルスケア」「スポーツ」の各 3 件のテキストを用いて解析を行った．結果として，カテゴリ「スポーツ」以外のテキストでは，カテゴリ推測 1 位の結果と本来属するカテゴリが一致した結果を得ることができた．

あるテキストに含まれる全ての専門用語の意味は，そのテキストがどのカテゴリに属するものか推測された結果から，そのカテゴリに属する Web 辞典を用いて表示される．そこで，テキストに含まれる専門用語に対して，そのカテゴリに含まれる Web 辞典によって適切な意味を表示できるのかを適合率と再現率を用いて評価する．カテゴリ「株（コンピュータ）」「コンピュータ」「ヘルスケア」から各 1 件のテキスト（ここではそれぞれ BU，CO，HE とする）を選出する．各テキストで取得できた用語の総数より，テキストにおいて専門的な意味を持つ全用語（以下，正解データ）を主観的に求めた．カテゴリ推測 1 位のカテゴリに属する Web 辞典を用いて，意味を調べられる正解データについて，適合率と再現率を求めた結果を表 4 に示す．

表 4: カテゴリ推測の適合率と再現率

カテゴリ (テキスト)	適合率	再現率
ビジネス (BU)	0.57	0.53
コンピュータ (CO)	0.89	0.60
ヘルスケア (HE)	0.73	0.77

4.2 表示された用語の意味の精度評価

カテゴリ「株（コンピュータ）」「コンピュータ」「ヘルスケア」に属するテキスト BU，CO，HE において，出現回数上位 20 単語に対して表示する意味の適合率について評価した結果を表 5 に示す．

表 5: 用語の意味表示における意味適合率

カテゴリ (テキスト)	意味適合率
ビジネス (BU)	0.85
コンピュータ (CO)	0.80
ヘルスケア (HE)	0.80

3 つのカテゴリにおいて，80% を超える精度でテキストに適した用語の意味表示を行える事が判明した．また，誤った意味表示をした用語については，以下の特徴が見られた．

- 2 番目の Web 辞典の方が正しい意味を持つ

- 複合名詞の一部の名詞が出現名詞として扱われる
- Web 辞典の追加説明部分に正しい意味が記載されている

5 考察

5.1 カテゴリ推測の精度について

テキストのカテゴリ「スポーツ」のみがシステムのカテゴリ推測において，本来のカテゴリと一致しない結果が得られた．これは利用したスポーツに関する各テキストにおいて，カテゴリ「スポーツ」としての特徴量が他のカテゴリ類よりも低いということを意味する．この事から，本システムは専門的なテキストであり，専門用語を多く含むテキストにおいて高い精度でカテゴリ推測が行えるものと考えられる．

本システムでは，適合率と再現率の観点から考えると，再現率が高いことが望ましい．何故なら，テキストに対して適した意味を持つ名詞には数に限りがあり，それらの名詞がカテゴリ推測の結果で 1 位に挙げられたカテゴリに属するのならば，本システムの手法において最適な意味を取得できる可能性が高くなると考えられるからである．しかし，表 4 より再現率にばらつきがある事が分かる．この事から，人が認識する専門的意味を持つ用語と Web 辞典が対応する専門用語の差に原因があると考えられる．

また適合率では，カテゴリ推測における正解率の信頼度は高いことから，Web 辞典が持つ有利性を本システムで利用できた結果であると考えられる．

5.2 表示された用語の意味の精度について

本システムの用語意味表示の精度として 80% を超える結果が得られた．また本システム評価で用いたテキスト BU，CO，HE について，各テキストの用語の意味を持つ Web 辞典の数について，以下の表 6 に示す．

表 6: 各テキストの用語の意味を持つ Web 辞典数

	用語に対応する Web 辞典数		
テキスト	平均	最大	最小
BU	3.35	5	1
CO	3.95	7	2
HE	4.75	10	1

表 6 より本システムでは，対応する Web 辞典を複数持つ用語に対して，Web 辞典を選定し適切な意味表示が行えている．以上より，本システムにおける手法において，高い信頼性や有用性があると考えられる．

しかし，カテゴリ推測の結果のみを利用して Web 辞典を選定しているため，用語の意味表示の際に，1 番目よりも 2 番目に選出した Web 辞典を利用する方が正しい意味である場合も存在した．また，Web 辞典の追加説明部分に正しい意味が記載されていた事も判明した．これは，カテゴリ推測の情報のみを用いた選定手法に原因があると考えられる．精度を高めるため，Web 辞典間での選定も行うべきである．

加えて，複合名詞の取得手法の見直しも検討すべきである．本システムでは，テキスト中から複合名詞の一部として取得された名詞についても，カテゴリ推測を実施する際に要素の一つとして用いている．ことから，カテゴリ推測に誤りが生じている可能性が考えられる．

6 おわりに

本研究では，テキストの分野に応じた用語説明システムの構築手法について提案した．シミュレーションの結果，本来属するカテゴリが抽出できたテキストに対し，約 80% の精度で適切な意味を出力できた．

今後の課題として，カテゴリ推測の結果に加えて Web 辞典間での優劣情報を用いた，Web 辞典の選定を行う必要がある．また，より多くの Web 辞典に対応させる事に加えて，カテゴリ推測時に用いられる名詞と複合名詞の扱い方を見直すことが挙げられる．

参考文献

[1] 三條場旭彦，藤井敦，“多面的な用語説明を生成するためのテキスト分類手法”，言語処理学会第 15 回年次大会発表論文集，pp.388-391，2009.

[2] 村脇有吾，黒橋禎，“テキストから自動獲得した名詞の分類”，言語処理学会第 16 回年次大会発表論文集，pp.716-719，2010.

[3] 藤井敦，“特許情報を専門用語の知識源として活用するシステム”，言語処理学会第 14 回年次大会発表論文集，pp.588-591，2008.