日本語語彙知識の統一的・整合的管理のデザイン

黒橋 禎夫　進 義治　柴田 知秀　村脇 有吾　河原 大輔
京都大学大学院情報学研究科
{kuro, shibata, murawaki, dk}@i.kyoto-u.ac.jp
shin@nlp.ist.i.kyoto-u.ac.jp

1　はじめに

自然言語の意味解析の研究を進める上で語彙知識が重要な役割を果たす．語彙知識には，形態素解析用辞書，国語辞典，シソーラス，格フレーム，百科事典など，さまざまなものがある．これまでの自然言語処理研究の蓄積と，特に集合知によるウェブ上の辞書によって，豊富な語彙知識が利用可能となったものの，現状の言語解析システムはこれらの情報を十分に使いこなしているとは言い難い．

従来の言語解析システムでは，これらの語彙知識を，適当なフォーマット変換を行って，適当な段階で，適当な表記のマッチングによって利用することが一般的であった．この方法では，余分な開発コストがかかることはもちろん，より本質的な問題として，単語間に区切りがなく，表記バリエーションの多い日本語においてはテキスト中の語彙の認識に失敗するという問題が発生する．

語彙知識を統一的に扱う枠組みの提案はセマンティックウェブのコミュニティーにおいて行われていたが ([1], [5])，日本語における単語区切りの問題に注意したものではないため，自然言語解析で利用する際には結局のところナイーブなマッチングを行うしかなかった．

そこで，本研究では，

- 日本語に関する様々な語彙知識を統一的に『語彙データベース』として管理し，
- 日本語文の最も基本的な解析，すなわち形態素解析と句認識の結果に語彙知識を埋め込む枠組みをデザインし，
- これを JUMAN，KNP の上で実装した．

この際，日本語における単語区切りが自然言語処理全体においても，語彙知識付与の上でも非常に重要であることに注意し，語彙知識全体を整合的な単語区切りで管理し，また同時に語彙知識から適切な形態素解析用辞書を自動生成することを実現した．これによって，日本語文の基本解析の一部として語彙知識が適切に埋め込まれるようになり，意味解析研究の障害の一つを取り除くことができたと考えている．

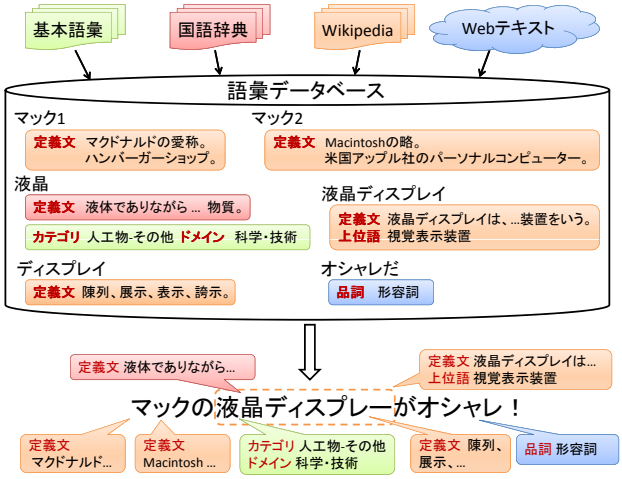

図 1: 語彙データベースによるテキストへの知識付与

2　語彙知識を扱う枠組み

2.1　単語区切りの整合性

日本語には単語区切りがなく，表記バリエーションが多いため，様々な語彙辞書等を利用する場合，見出し表記を単純に用いるだけでは問題が生じる．たとえば,「コンピュータグラフィックス」に関する語彙知識がこの表記の見出し語で与えられている場合，単純なマッチングではテキスト中の「コンピューターグラフィックス」「コンピュータグラフィクス」などにマッチしない．これはカタカナ表記に限らず，かな漢字，漢字異表記を含め日本語語彙に遍在する問題である．

我々は，形態素解析システム JUMAN において，基本語 3 万語を策定し，その表記バリエーションに対し

```
<LDEntry id="1605558" type="Wikipedia">
  <RepForm>液晶/えきしょう+ディスプレー/でぃすぷれー</RepForm>
  <Form>液晶ディスプレイ</Form>
  <POS>名詞</POS>
  <Def>液晶ディスプレイは、液晶組成物を利用する平面状で薄型の
       視覚表示装置をいう。… </Def>
  <Sem type="上位語">視覚表示装置</Sem>
</LDEntry>
```

RepForm : 代表表記 , Form : 表記, POS : 品詞, Def : 定義文,
Sem : 意味情報

図 2: 語彙知識のエントリの例

て代表表記というある種の ID を付与するという枠組みを提案している．ここではこれを拡張し，複合語も含めて語彙表記全体を整合的に代表表記で扱うこととした．例えば,「コンピュータグラフィックス」は

> コンピューター/こんぴゅーたー+グラフィックス/ぐらふぃっくす

のように，単語の代表表記を “+” でつなげた形で表現する．ここで,「コンピューター/こんぴゅーたー」は「コンピューター」「コンピュータ」の代表表記であり，これによって上述のような問題を解消することができる (言語解析システムの実際の動きについては 2.3 節で述べる).

2.2 語彙データベース：フォーマットの統一

様々な知識源から獲得した知識を，統一的なフォーマットで記述し統合したデータベースを構築する．このデータベースを「語彙データベース」と名付ける．現在，語彙データベースに入れている知識源の一覧を表 1 に示す.

語彙データベース中の各語彙知識は図 2 のような XML のエントリを単位として記述する．なお，多義語は，語義ごとに持っている意味情報が異なるため，語義ごとに異なるエントリを作成する.

2.3 語彙データベースに基づく言語解析システム

語彙データベースの情報（インデックス）の付与は，形態素解析や句認識の後に改めて行うのではなく，形態素解析や句認識と同時に行う方が，効率面からも利用者の利便性からも望ましい.

まず，そもそも形態素解析用の語彙情報も語彙データベース中で管理しているので，ここから形態素解析器 JUMAN 用の辞書を自動生成する．単語に関する他

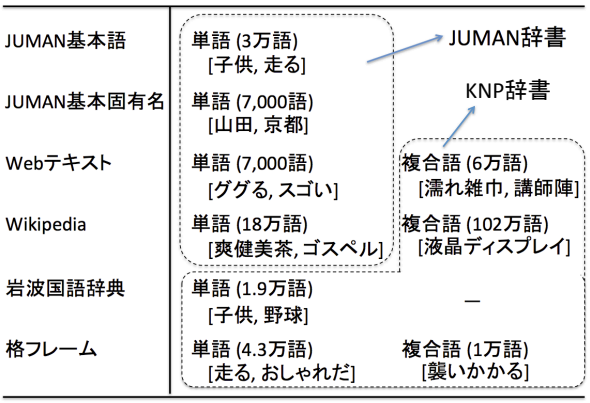

表 1: 様々な知識源

の語彙情報と，複合語に関する語彙情報は，構文解析器 KNP 用の辞書の形で自動生成し，KNP の中の最初の処理である句認識においてその語彙情報を付与する．なお，この句認識において，複合語の語彙情報 (3.4.2 節) に基づいて品詞を修正する処理も行っている.

3 語彙データベースの知識源

我々は現在，基本語彙，国語辞典，Wikipedia，Web テキストの 4 種類の知識源から語彙知識を獲得し，語彙データベースを構築している．各知識源について，獲得する知識の種類および獲得手法を示す.

3.1 基本語彙

形態素解析器 JUMAN に付属していた基本語と基本固有名の形態素解析辞書を，語彙データベースの知識源として扱う．JUMAN では基本語 3 万語を人手で選定しており，それぞれに意味情報を付与している．また，JUMAN では基本語とは別に，人名，地名からなる 7,000 語の基本固有名辞書を整備している.

3.2 国語辞典

定義文は語の意味情報として基本的なものであり，様々な用途で利用できる.

複合語は 3.3 節で述べる Wikipedia に情報があり，用言については 3.4.4 節で述べる格フレームを利用するため，ここでは，JUMAN 基本語に対して定義文の情報を付加したエントリを生成する．多義語はそれぞれの語義に対応する定義文が別々に記述されているため，語義ごとに一エントリを生成する.

国語辞典として岩波国語辞典を用いているが、将来的には Wiktionary を利用することを検討している.

3.3 Wikipedia

Wikipedia は幅広いドメインの語をカバーしており，様々な語彙知識を獲得することができる．

Wikipedia の各記事から見出し語 (記事のタイトル) と先頭段落を抽出し，一エントリを生成する．また，先頭段落の一文目から上位語を抽出する．例えば,「液晶ディスプレイは，液晶組成物を利用する平面状で薄型の視覚表示装置をいう.」からは上位語「視覚表示装置」を抽出する．

通常の記事のほかに曖昧さ回避ページとリダイレクトページがある．曖昧さ回避ページは，複数の意味がある語に対して意味を一覧表示し，ユーザの調べたい意味へ誘導する役割を果たす．曖昧さ回避ページを持つ語は多義語とみなせ，意味の一覧から各語義を得ることができる．各語義ごとに一エントリを生成し，例えば，多義語「マック」に対してはマクドナルドやマッキントッシュを表すエントリを生成する．また，リダイレクトページは，別のページへ転送する役割を果たし，略称や別名で記事を検索しても実際の記事へ転送する．リダイレクトページからは同義語に関する知識が獲得できる．例えば,「京大」は「京都大学」にリダイレクトされており，ここから「京大」のエントリを作成し，意味情報に「リダイレクト:京都大学」を付与する．

見出し語は「子供」のような JUMAN 基本語に含まれる語,「爽健美茶」や「ミニストップ」のような JUMAN 基本語にはない単語,「京都大学」のような複合語が含まれる．JUMAN 基本語にはない単語の認識については，基本語の辞書を用いた形態素解析結果に基づき解析誤りと思われるものを一単語であると判断する [3]．2.3 節で述べたように，これらの語から形態素解析用の辞書を生成する．

3.4 Web テキスト

3.4.1 形態素

未知語による形態素解析の誤りを解消するため，形態素の自動獲得を行う．3.3 節で述べた通り，我々は Wikipedia から形態素の獲得を行っているが，Wikipedia の見出し語には用言があまり含まれていない．そこで，Web テキストからは,「ググる」,「カサつく」などの用言を獲得する [2].

3.4.2 複合語

「蒸し鶏」や「反面教師」のように，動詞連用形+名詞，副詞+名詞で構成される複合語があり，構文解析においてこれらの認識が重要である．

これらは Wikipedia のエントリとして存在する場合があるが,「濡れ雑巾」など存在しない表現も大量にあるため，Web テキストから自動的に獲得している．これは，テキスト中の連続する二語について，自己相互情報量 (PMI) が閾値より高いものを抽出することによって行っている．

3.4.3 複合語カテゴリの自動推定

JUMAN 基本語には「人」,「場所-施設」などのカテゴリが付与されている．複合語のカテゴリは基本的には主辞のカテゴリとみなせばよい．しかし，複合語の主辞が漢字一文字の場合は複合語のカテゴリとして採用すると誤りとなる場合がある．例えば,「講師陣」の場合,「陣」のカテゴリは「場所-その他」または「抽象物」であり，これを「講師陣」のカテゴリとするのは誤りとなる．

そこで，末尾が一形態素の複合語に対して，カテゴリが付与された基本語すべてと分布類似度を計算し，k 近傍法で複合語のカテゴリを推定する．例えば,「講師陣」の場合，基本語「スタッフ」,「メンバー」「講師」などと高い類似度をとり，これらのカテゴリは「人」であることから,「講師陣」のカテゴリは「人」と推定される．

3.4.4 格フレーム

格フレームとは，用言とその格要素の間の関係知識を記述したものである．我々はこれまでに大規模テキストから格フレームを構築する手法を提案しており [4]，構文解析や省略解析で用いている．

「経験を積む」と「荷物を積む」のように，同じ表記の用言であっても語義が異なる場合は別の格フレームとして構築されており，エントリは格フレームごとに生成する．

4 解析例

図 3 に文「マックの液晶ディスプレーがオシャレ！」に対する KNP の解析例を示す．この例では Wikipedia や岩波などのエントリが埋め込まれている.「<LD- ...>が 1 エントリを表し，タグ内は属性と属性値のペア

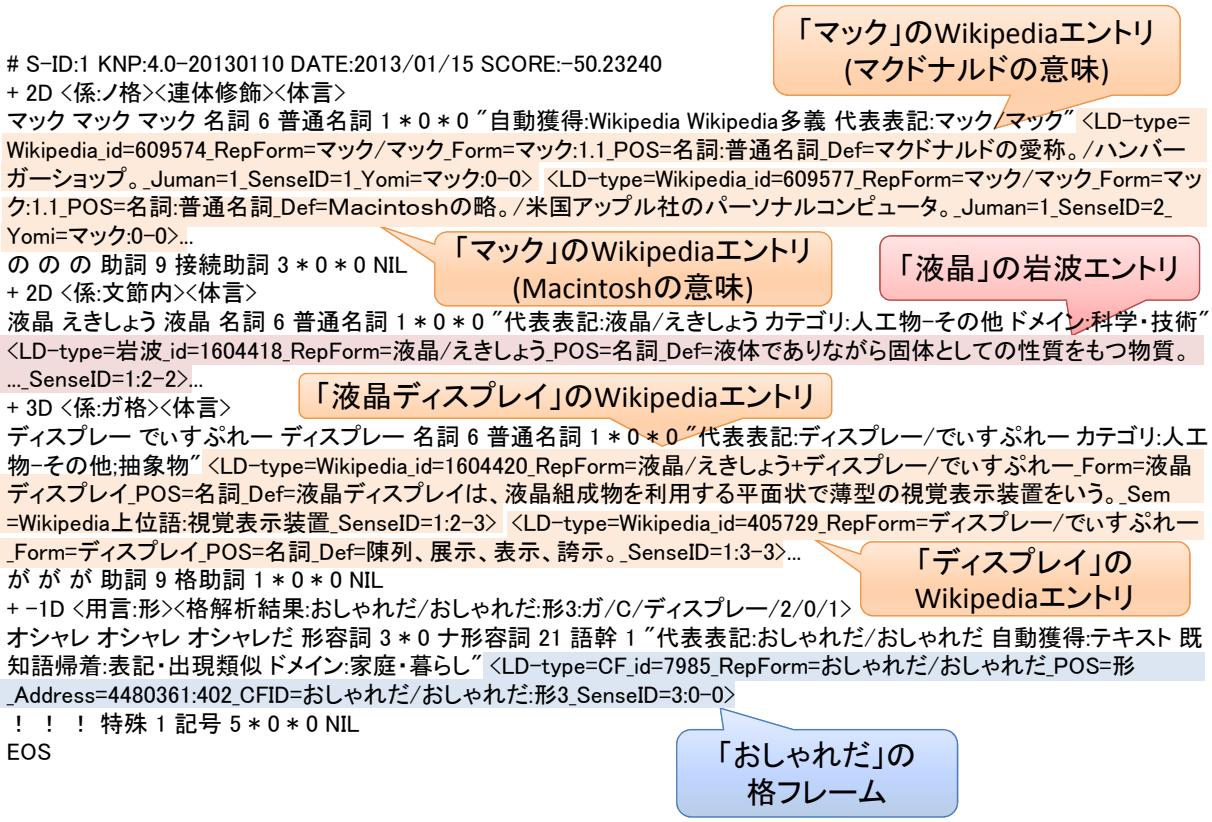

図 3: 文「マックの液晶ディスプレーがオシャレ！」の解析例

の集合からなる．代表表記でマッチングを行なうことにより，Wikipedia の「液晶ディスプレイ」と入力文中の「液晶ディスプレー」をマッチすることができている．

KNP での処理の最初の方で語彙知識を付与しているため，KNP で行っている構文解析や省略解析などの解析で語彙知識を利用することが可能となり，また，情報検索や機械翻訳などのアプリケーションにおいて，KNP の解析結果を読むだけで様々な語彙知識を利用することが可能となる．

5 おわりに

本論文では，様々な知識源から獲得した語彙知識を統一的なフォーマットで記述して統合的に扱う枠組みを構築し，その枠組みに基づいて言語解析システムで語彙知識を利用するための手法を提案した．

語彙知識を言語解析で利用することを始めており，3.4.3 で説明した，「講師陣」のカテゴリが「人」であることを利用し，省略解析の誤りが改善するなどの例を確認している．構文解析や省略解析などの言語解析において大規模に評価し，語彙知識を利用することの有用性を示すのは今後の課題である．また，名詞関連語などの情報を語彙知識に追加することや，それを用いて多義性解消を行う予定である．

参考文献

[1] DBpedia. http://dbpedia.org/.

[2] Yugo Murawaki and Sadao Kurohashi. Online acquisition of Japanese unknown morphemes using morphological constraints. In *Proceedings of EMNLP2008, poster*, pp. 429–437, 2008.

[3] 柴田知秀, 村脇有吾, 黒橋禎夫, 河原大輔. 実テキスト解析をささえる語彙知識の自動獲得. 言語処理学会 第 18 回年次大会, pp. 81–84, 2012.

[4] 河原大輔, 黒橋禎夫. 格フレーム辞書の漸次的自動構築. 自然言語処理, Vol. 12, No. 2, pp. 109–131, 2005.

[5] 武田英明. 日本における linked data の現状と普及に向けた課題. 情報処理, Vol. 53, No. 3, pp. 326–333, 2011.