社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－ＦＩ－ 76　(7)
2004－ＮＬ－163　(7)
2004／9／16

新聞記事からの用語定義の抽出と固有表現クラスに基づく分類

小山　誠　　酒井　哲也　　真鍋　俊彦

(株) 東芝 研究開発センター 知識メディアラボラトリー
〒212-8582 川崎市幸区小向東芝町 1
{makoto3.koyama, tetsuya.sakai, toshihiko.manabe}@toshiba.co.jp

あらまし　本報告では，質問応答システムなどの自然言語処理システムの言語知識の拡張のため，新聞記事から用語定義を抽出し，分類・体系化するシステムを提案する．本システムは，定義文に対する固有表現抽出結果から得られる固有表現の意味クラスと，定義文に対する形態素解析結果から抽出される語に基づき，用語定義を分類する．新聞記事を用いた評価実験を行った結果，14 の意味クラスに対して，適合率 82.1%，再現率 50.8%で抽出した用語定義を分類できることを確認した．

Extraction and Classification of Term Definitions Using Named Entity Extraction from News Articles

Makoto KOYAMA　　Tetsuya SAKAI　　Toshihiko Manabe

Knowledge Media Laboratory, Corporate Research & Development Center,
TOSHIBA CORPORATION
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki 212-8596, Japan
{makoto3.koyama, tetsuya.sakai, toshihiko.manabe}@toshiba.co.jp

Abstract In this paper, we propose a system that uses Japanese newspaper corpora for extracting and classifying term definitions to expand the knowledge of a natural language system such as a question answering system. The system classifies term definitions based on semantic classes obtained through named entity extraction and words obtained through morphological analysis. In an experiment using news articles, the system classifies term definitions by 14 semantic classes and achieves 82.1% precision and 50.8% recall.

1 はじめに

近年，World Wide Web や新聞記事データなどの大量の情報を用いた質問応答システムの研究開発が盛んになってきている[1]．こうした質問応答システムにおいて，定義を問うタイプの質問（例えば，"ブルートゥースとは何？"などの質問）に回答することが新しい研究課題の一つになっている[2]．本研究は，こうした定義に関する質問に回答するための，用語定義（用語とその定義）を集めた百科辞典的知識を構築することを目的とする．

定義に関する質問に回答する知識源として，人手によって編纂された国語辞典や百科事典などを利用することが考えられる．しかし，こうした辞典・事典情報は，新しい用語に弱い場合が多く，収録される語数も限られているため，知識として充分でない．

また一方，新聞記事などのテキストデータに含まれる定義や説明を利用することが考えられる．新聞記事において新しい用語や専門用語などが出現したときは，通常そうした用語の定義や説明も記事に含まれる．

こうした，記事に含まれる定義情報を収集し，それらを分類・体系化することにより，質問応答のための百科事典的知識を自動または半自動で構築することができると考えられる．

そこで本報告では，新聞記事データを対象に，用語の定義情報を抽出し，それらを質問応答システムの固有表現クラスに基づき分類・体系化するシステムを提案する．本システムは，用語とその定義情報をパターンマッチングによって抽出し，さらに抽出した用語をその定義文に対する固有表現抽出結果に基づき分類する．

以下，2 章でシステムの概要を述べた後，3 章で用語定義の収集・抽出，4 章で抽出した用語定義の分類について説明する．5 章では毎日新聞記事を用いた評価実験の結果について報告する．6 章で関連研究について述べ，最後に 7 章でまとめと今後の課題について述べる．

2 システムの概要

図 1 に提案するシステムの処理の流れを示す．まず，新聞記事データから，記事のタイトルの形式に基づき用語定義が含まれる記事を収集する．次に，収集した記事からパターンマッチングにより用語とその定義文，定義文を含む定義段落を抽出する．そして，抽出した定義文を形態素解析，固有表現抽出し，その結果に基づき用語を分類する．ここで蓄積された知識は，質問応答システムの言語知識として利用される．

3 記事の収集と用語定義の抽出

3.1 記事の収集

新聞記事データの中から用語定義が含まれる記事を収集する．新聞記事データには用語の説明を主題とする記事があり，こうした記事の多くはタイトルが定形のフォーマットで記述されている．図 2 に記事のタイトルの例を示す．こうした形式のタイトルに対応するパターンを人手で作成し，パターンマッチングにより新聞記事データ中から用語定義が含まれる記事を収集する．

本研究では，毎日新聞 98 年版，99 年版，読売新聞 98 年版，99 年版の記事のタイトルを調べてパターンを作成した．

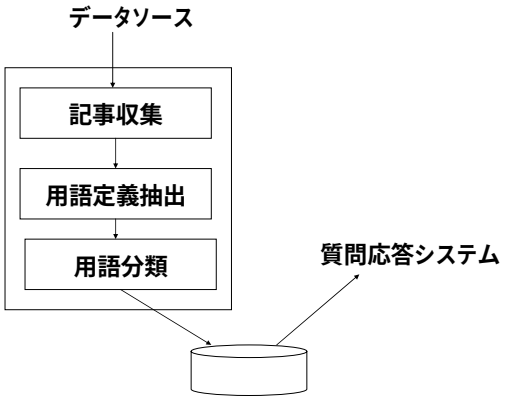

図 1 システムの処理の流れ

［用語］ＢＯＣ
［特集］デジタル時代を読む１９９８　今週のキーワード　ＥＲＰ
［ニュースの言葉］ＤＳＬ

図 2 記事のタイトルの例

3.2 用語定義の抽出

人手で作成したパターンを用いて、収集した記事から用語とその定義を抽出する．まず，用語の抽出から行う．ここでは，記事タイトルにおける用語と，記事本文中の用語見出しにおける用語を抽出対象とする．用語見出しは，記事本文において例えば，“◆　ＡＢＣ”のように表される記述であり，この見出しの後に用語「ＡＢＣ」の説明が記述される．記事タイトル，用語見出しそれぞれから用語を抽出するパターンを作成し，パターンマッチングにより抽出を行う．

次に，抽出した用語に対する定義文とその定義文を含む定義段落を抽出する．定義文は次のような，用語自身を含まない定義文と用語を含む定義文とがある．

- 用語を含まない定義文の例
  「…な装置。」
  「…方式の一つ。」

- 用語を含む定義文の例
  「Ａとは…」
  「Ａは…」
  「…をＡという」

@DISEASE→@DISEASE
@PRODUCT→@PRODUCT
疾患→@DISEASE
装置→@PRODUCT

図 3 分類ルールの例

タイトルの次の文（本文 1 文目），または用語見出しの次の文に書かれている場合には，用語自身を含まない形式で記述されることが多くなっている．一方，タイトル，用語見出しの次の文に定義文がない場合は，本文中に用語自身を含む形式で記述されることが多くなっている．それぞれに対する抽出のパターンを作成し，次のようにして用語に対する定義文の抽出を行う．

(1) 記事本文の 1 文目または用語見出しの次の行から，パターンマッチングにより，用語が含まれない定義文を抽出．

(2) (1)で定義文が抽出されなければ，本文中から，パターンマッチングにより，用語が含まれる定義文を抽出．

最後に，抽出した定義文を含む定義段落を求める．ここでは経験的に定義文の後の 3 文を抽出し，それを定義段落とした．

4 用語の分類

抽出した用語を，その定義文に対する固有表現抽出結果と形態素解析結果に基づき分類する．

4.1 意味クラス，上位語による分類

用語と定義文を固有表現抽出，形態素解析し，その結果から，定義に含まれる固有表現の意味クラス，および，用語の上位語を求める．意味クラス，上位語は，それぞれ定義文文末，括弧付き表現，用語末尾から抽出する．それぞれにおける固有表現抽出結果と形態素解析結果の例を以下に示す．ここで，@A は抽出された固有表現の意味クラスを表す．

- 定義文文末
  “ …を原因とする**@DISEASE**。”
  “ …/を/原因/と/する/**肝炎**/。/ ”

- 括弧付き表現
  (a) 括弧の外
  “ …**@RULE**（ＮＰＴ）は、… ”
  “ …/**条約**/（/ＮＰＴ/）/は/、/… ”
  (b) 括弧の中
  “ …NPT(**@RULE**)は、… ”
  “ …/NPT/(/…/**条約**/)/は/、/… ”

- 用語末尾
  “ /ページ/**システム**/ ”

定義文文末からの意味クラス，上位語の抽出については，前処理で特定の文字列を取り除いた後，末尾の意味クラスまたは名詞を抽出する．例えば，“ …/方式/の/一つ/．/ ”などのように記述されている場合は，前処理で “ 一つ ” を除いた後，“ 方式 ” を上位語として取り出す．

括弧付き表現からの抽出は，括弧記号を挟んで用語と隣接する語から意味クラスまたは名詞（複合語の場合は最後尾の名詞）を抽出する．

また，用語が複合語となっている場合は，その最後尾の名詞を上位語として抽出する．

以上のようにして抽出した意味クラスと上位語に基づき，用語を分類する．このとき，図 3 に示すようなルールを作成して分類する．ルールは，意味クラスを用いたルール（“ @PRODUCT→@PRODUCT ” など）と上位語を用いたルール（“ 装置→@PRODUCT ” など）を作成する．“ @PRODUCT→@PRODUCT ” は，定義文から@PRODUCT の固有表現が抽出されたときに，用語を@PRODUCT に分類するルールを表す．“ 装置→@PRODUCT ” は定義文から上位語として “ 装置 ”が抽出されたときに，用語を@PRODUCT に分類するルールを表す．

4.2 上位語の多義性解消

上位語による分類では，抽出される上位語が多義語となっている場合がある．例えば，“ システム ” という語は@RULE の語の上位語になる場合もあり，また@PRODUCT の語の上位語になる場合もある．上位語

がこのように多義語となる場合, 複数のルールが作られる. 先の“システム”についていえば, @RULE に分類するルール(“システム→@RULE”)と @PRODUCT に分類するルール(“システム→@PRODUCT”)の 2 つが作られる.

ここで, 定義文に“システム”を含む新たな用語を分類するとき, 複数の分類先から正しい分類先を決定する必要がある. そこで, 定義文とその後の 3 文までを含む段落を文脈として, 上位語の多義性の解消を行うことにより, 正しい分類先を決定する. この多義性の解消を含めた用語の分類方法は次のようになる.

(1) ルールにより分類先が 1 つになる用語を分類する.

(2) (1)で分類された用語の段落を形態素解析し, 段落に含まれる各単語の頻度を計算する.

(3) ルールにより分類先が複数ある用語について, 上位語の多義性の解消を行い, 分類先を決定する. ここで, 多義性の解消には, [3]で提案されている手法を用いた. 次式により分類先を決定する.

$$\arg\max_{c} \sum_{w \in context} \log \frac{\Pr(w \mid c) \times \Pr(c)}{\Pr(w)}$$

$\Pr(w)$：単語 w の出現確率
$\Pr(c)$：意味クラス c の出現確率
$\Pr(w \mid c)$：c における w の出現確率

ここで, $\Pr(w)$, $\Pr(w \mid c)$は(2)で各意味クラスに分類された用語の定義段落に含まれる単語により計算した. $\Pr(c)$は定数とした. また, *context* は, 対象となる用語の定義段落とする.

5 評価実験

新聞記事データを用いて用語定義の抽出と分類の評価実験を行った. 固有表現抽出には, 我々のグループが開発している質問応答システム[4]における固有表現抽出システム[5]を利用した.

表 1 各意味クラスの用語数

意味クラス名	説明	数
RULE	法律名、規則名、制度名	280
PRODUCT	製品カテゴリー、製品名	170
ORGANIZATION	組織名	115
SUBSTANCE	物質名	84
DISEASE	病名、傷名、症状名	52
CONSTRUCTION	建造物名	47
ANIMATE	生物名	43
UNIT	単位名	41
LOCATION	場所名	26
PERSONAL_ATTRIBUTE	地位名、職業名	23
ISSUE	社会問題	20
EVENT	イベント名、会議名	19
TRANSPORT	交通手段カテゴリー、乗物名	17
WEAPON	武器カテゴリー、武器名	14

5.1 正解データの作成

新聞記事データとして毎日新聞記事データ 98 年版, 99 年版を使用し, ここから用語定義を含む記事を記事のタイトルに基づき収集した. 毎日新聞 98 年版からは 451 記事, 99 年版からは 458 記事を収集した. これらから人手により用語と, 定義が含まれる段落を抽出した. こうして抽出した段落の第 1 文を正解の定義文とした. さらに, 各用語を意味クラスに分類し, これを正解データとした. 評価には, 分類された用語数が多かった 14 の意味クラスを用いる. 表 1 に, これらの意味クラスと各意味クラスにおける用語数を示す.

5.2 抽出の評価

毎日新聞 98 年版の正解データに基づき人手により, 用語および定義文抽出のパターンを作成した. このルールを用いて, 毎日新聞 98 年版, 99 年版から収集した記事を対象に, 用語定義の抽出を行い, 抽出の有効性を評価した. 評価には, 適合率, 再現率, F 値を用いた. それぞれ次式で計算した.

表 2 定義抽出の結果

	適合率	再現率	F値
毎日新聞98	94.3(315/334)	69.8(315/451)	80.2
毎日新聞99	90.8(317/349)	69.2(317/458)	78.5

表 3 毎日新聞 98 の分類結果

	適合率	再現率	F値
多義性解消なし	85.9(286/333)	98.6(286/290)	91.8
多義性解消あり	95.6(281/294)	96.9(281/290)	96.2

表 4 毎日新聞 99 の分類結果

	適合率	再現率	F値
多義性解消なし	72.4(163/225)	53.1(163/307)	61.3
多義性解消あり	82.1(156/190)	50.8(156/307)	62.8

$$適合率 = \frac{システムが出力した正解数}{システムが出力した結果の数} \times 100$$

$$再現率 = \frac{システムが出力した正解数}{全ての正解} \times 100$$

$$F = \frac{2 \times 適合率 \times 再現率}{適合率 + 再現率}$$

ここで，正解の判定は，抽出された段落に正解の定義文が含まれているか否かを調べ，含まれていれば正解とした．

表 2 に毎日新聞 98 年版，99 年版に対する適合率，再現率，Ｆ値を示す．これより，適合率は毎日新聞 98 年版，99 年版ともに 90%を超えており，収集した記事から高い精度で用語定義を抽出できていることが分かる．一方，再現率は毎日新聞 98 年版，99 年版ともに約 69%になっている．これは，定義文の抽出が失敗している場合と，用語の抽出が失敗している場合とがあった．

- 定義文抽出の失敗

主に，上位語を含まない定義文の抽出が失敗していた．例えば，用語「ＰＣＢ」の定義文は次のように上位語のない形式で記述されている．

“耐熱，絶縁性に優れ、コンデンサーなどの絶縁体として使われてきた。“

本研究では，主に上位語を含む定義文のパターンを作成したが，再現率を上げるためには，こうした上位語を含まない定義文に対するパターンを充実化する必要がある．

- 用語抽出の失敗

本研究では，用語抽出のパターンとして，１つの用語見出しから 1 つの用語を抽出するパターンを作成している．新聞記事中には，例えば，“ＢＯＤとＣＯＤ”のような複数の用語が含まれる用語見出しもあるが，こうした見出しからは用語を抽出していない．また，抽出結果のノイズを減らすため，「４０１Ｋ」のような数字で始まる用語は抽出していない．再現率を上げるためには，こうした場合に対応するよう，抽出パターンのさらなる充実化と精緻化が必要である．

また，毎日新聞 98 年版の結果と 99 年版の結果とを比較すると，それらの結果は大きくは変わっていない．これより，98 年版で作成したルールでも，99 年版から精度を落とすことなく用語定義を抽出することが可能なことが分かる．

表 5 毎日新聞 99 の意味クラスごとの分類結果

	多義性解消なし			多義性解消あり		
意味クラス	適合率	再現率	F値	適合率	再現率	F値
@RULE	77.8(63/81)	56.8(63/111)	65.6	89.6(60/67)+	54.1(60/111)-	67.4+
@PRODUCT	60.0(21/35)	52.5(21/40)	56.0	714(20/28)+	50.0(20/40)-	58.8+
@ORGANIZATION	78.6(22/28)	75.9(22/29)	77.2	76.9(20/26)-	69.0(20/29)-	72.7-
@SUBSTANCE	66.7(10/15)	71.4(10/14)	69.0	64.3(9/14)-	64.3(9/14)-	64.3-
@DISEASE	93.8(15/16)	75.0(15/20)	83.3	100(15/15)+	75.0(15/20)	85.7+
@CONSTRUCTION	100(5/5)	20.0(5/25)	33.3	100(5/5)	20.0(5/25)	33.3
@ANIMATE	66.7(6/9)	50.0(6/12)	57.1	85.7(6/7)+	50.0(6/12)	63.2+
@UNIT	75.0(6/8)	42.9(6/14)	54.5	75.0(6/8)	42.9(6/14)	54.5
@LOCATION	50.0(3/6)	25.0(3/12)	33.3	75.0(3/4)+	25.0(3/12)	37.5+
@PERSONAL_ATTRIBUTE	66.7(2/3)	33.3(2/6)	44.4	66.7(2/3)	33.3(2/6)	44.4
@ISSUE	0.00(0/6)	0.00(0/1)	0.0	0.00(0/0)	0.00(0/1)	0.0
@EVENT	66.7(2/3)	28.6(2/7)	40.0	66.7(2/3)	28.6(2/7)	40.0
@TRANSPORT	71.4(5/7)	45.5(5/11)	55.6	71.4(5/7)	45.5(5/11)	55.6
@WEAPON	100(3/3)	60.0(3/5)	75.0	100(3/3)	60.0(3/5)	75.0
計	72.4(163/225)	53.1(163/307)	61.3	82.1(156/190)	50.8(156/307)	62.8

5.3 分類の評価

毎日新聞 98 年版の正解データを利用して，5.1 節の 14 クラスへの分類ルールを作成した．まず，用語とその定義文を形態素解析し，その結果から上位語を抽出した．それらの上位語について，各意味クラスへのルールを作成した．また，固有表現システムにおける各意味クラスから分類先となる 14 の意味クラスへの対応を作成した(“@MAMMAL→@ANIMATE”など).

こうして，作成したルールを用いて，毎日新聞 98 年版，99 年版の用語を分類した．入力は自動抽出した用語とその定義である．

分類結果の評価には適合率，再現率，F 値を 5.2 節の式で計算した．ここで，正解はシステムの出力した分類と人手で付けた分類が一致した場合とした．

表 3 に毎日新聞 98 年版における用語の分類結果の適合率，再現率，F 値を示す．また，表 4 に毎日新聞 99 年版における用語の分類結果の適合率，再現率，F 値を示す.

表 3, 表 4 より，多義性の解消により，毎日新聞 98 年版では 9.7 ポイント，毎日新聞 99 年版では 9.7 ポイント適合率が上昇している．一方，再現率はそれぞれ，1.7，2.3 ポイントの低下となっており，再現率の低下は低く抑えられている．

本実験では，分類のルール作成には，毎日新聞 98 年版の定義文に出現した固有表現の意味クラスおよび上位語のみを用いた．表 4 より，これらのルールだけで，毎日新聞 99 年版の用語の約 50%は分類できている．しかし，表 3 の毎日新聞 98 年版の結果と比較すると再現率は大きく落ちている．これは，毎日新聞 98 年版から抽出される上位語だけでは，毎日新聞 99 年版の用語を分類するのに十分なルールが作成されなかったためである．分類が失敗した多くの用語については，上位語として正しいと思われる単語が獲得されている（例えば，@RULE の用語では“規則”・“条項”・“新法”など，@PRODUCT の用語では“ロボット”・“オーディオ”・“端末”など，@ANIMATE の用語では“昆虫”・“猿人”・“マグロ”などが獲得されている）が，対応するルールが無いため，正しく分類されていなかった．再現率を上げるためには，シソーラスなどの言語知識を利用して，上位語に関するルールを強化する必要があると考えられる．

表 5 に毎日新聞 99 年版のクラスごとの分類結果の適合率，再現率，F 値を示す．表の中で，多義性の解消によりそれぞれの数値が上がった場合には“＋”を，また下がった場合には“－”を付けている．特に，@RULE, @PRODUCT などで適合率が上昇している．これらのクラスでは“システム”など語の多義性が解消されたためである．

多義性の解消により，分類結果の F 値が変化したものは 24 件あった．このうち，17 件は F 値が上がり，7 件は値が下がっており，符号検定（α=0.05）により有意な差がみられた．しかし，データ数は多くないため，他のデータを用いた評価を行うなど，今後さらに有効性の検証を行う必要がある．

6 関連研究

World Wide Web や新聞記事などの情報中から用語の定義や説明を抽出し, それらを分類またはグループ化している研究には[6], [7], [8]などがある.

[6], [7]は, World Wide Web を対象に用語説明の抽出を行っている. [6]は, 抽出した用語説明を, 文書分類手法を利用して 19 の専門分野に分類している. [7]は, 抽出した用語説明を, 定義種別により分類している. また, 同じ内容の用語説明どうしをグループ化し, そこから代表的な用語説明と上位語の抽出を行っている. [8]は, 新聞記事を対象に用語の説明文を抽出して用語集の作成を試みている. この研究では, 用語とその説明文を, 用語と説明文との間の意味関係（Is_a など）に基づき分類している.

これらの研究に対して, 本研究は, 新聞記事から抽出した用語とその定義を, 用語の意味クラスに基づき分類している

7 おわりに

新聞記事から用語定義を抽出し, 固有表現の意味クラスに基づき分類するシステムを提案した. システムの評価として新聞記事データを用いた抽出・分類実験を行い, 有効性と課題を検証した. 14 の意味クラスに対して, 適合率 82.1%, 再現率 50.8%で抽出した用語定義を分類できることを確認した. 今後の課題としては,

- 用語抽出, 定義文抽出パターンの充実化・精緻化による抽出性能の改善
- シソーラスなどの言語知識の利用による分類性能の改善
- 他のデータを用いた実験

などが挙げられる.

参考文献

[1] Xu, J. et al.: Evaluation of an Extraction-Based Approach to Answering Definitional Questions, *Proceedings of ACM SIGIR 2004*, pp.418-424, 2004.

[2] Voorhees, E. M.: Overview of the TREC 2003 Question Answering Track, *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, 2003.

[3] Yarowsky, D.: Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora, *Proceedings of COLING-92,* pp.454-460, 1992.

[4] Sakai, T. et al.: ASKMi: A Japanese Question Answering System based on Semantic Role Analysis, *Proceedings of RIAO2004*, pp.215-231, 2004.

[5] 市村ほか, 質問応答と, 日本語固有表現抽出および固有表現体系の関係についての考察, 情処学会研究報告, NL161-3, 2004.

[6] 藤井ほか, World Wide Web を用いた事典知識情報の抽出と組織化, 電子情報通信学会論文誌 D-Ⅱ, Vol.J85-D-Ⅱ, No.2, pp.300-307, 2002.

[7] 桜井ほか, ワールドワイドウェブを利用した用語説明の自動生成, 情報処理学会論文誌, Vol.43, No.5, pp.1470-1480, 2002.

[8] 山田ほか, ニュース記事に出現する用語と説明文の意味関係自動獲得, 情処学会研究報告, NL152-21, 2002.