社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

決定リストを用いた語義曖昧性解消

八木　豊† 野呂　智哉† 白井　清昭†† 徳永　健伸† 田中　穂積†

† 東京工業大学
†† 北陸先端科学技術大学院大学

E-mail: †{yutaka,noro,take,tanaka}@cl.cs.titech.ac.jp, ††kshirai@jaist.ac.jp

あらまし　本論文では，語義曖昧性解消のコンテスト SENSEVAL2 日本語辞書タスクの枠組において，決定リストを用いた手法について述べる．辞書タスクの訓練データには形態素情報と文書分類を表す UDC コードが与えられている．我々は既存のパーザを使って構文情報を追加し，合計 3 つの情報から得られる証拠をもとに決定リストを作成した．また，訓練データ不足を補うために，岩波国語辞典の語釈文中に含まれる例文を利用した．その結果，79.1%の正解率で語義曖昧性解消を行うことができた．さらに論文中では，どの情報から得られる証拠が有効であったかについても考察する．

キーワード　SENSEVAL2，辞書タスク，決定リスト，語釈文，国際十進分類法

Decision lists for Japanese word sense disambiguation

Yutaka YAGI†, Tomoya NORO†, Kiyoaki SHIRAI††, Takenobu TOKUNAGA†, and Hozumi TANAKA†

† Tokyo Institute of Technology
†† Japan Advanced Institute of Science and Technology

E-mail: †{yutaka,noro,take,tanaka}@cl.cs.titech.ac.jp, ††kshirai@jaist.ac.jp

Abstract　In this paper, we describe Japanese word sense disambiguation in the monolingual dictionary-based task of the SENSEVAL2 framework. Our system is based on decision lists. To create decision lists, we extracted evidences from three linguistic information, which are morphological information, syntactical information and UDC code. For 79.1% of target words in evaluation data, the appropriate word senses were assigned. Furthermore, we investigate what type of evidence is useful for word sense disambiguation.

Key words　SENSEVAL2, dictionary-based task, decision list, definition sentense, Universal Decimal Classification

1. はじめに

2001年の4月から5月にかけて，語義曖昧性解消のコンテストSENSEVAL2が行われた．SENSEVAL2では各言語ごとにいくつかのタスクに分かれており，日本語には2つのタスクがある．ひとつは日本語から英語への訳語選択を行う翻訳タスク，もうひとつは岩波国語辞典の語釈の中から適切な語釈を選ぶ辞書タスクである．我々は，決定リスト [1] を用いて語義曖昧性解消を行うシステムを作成し，辞書タスクに参加した．

決定リストはクラス分類を行う機械学習手法の1つである．自然言語処理における問題の多くはクラス分類の問題として捉えることができることから，近年では，アクセント記号付与 [6]，単語分割 [3]，固有表現抽出 [2]，形容詞の修飾先の判定 [4]，語義曖昧性解消 [7] など，様々な問題に決定リストを応用した研究が報告されており，その有効性が示されている．特に，前回行われたSENSEVAL1では，決定リストを階層的に適用するように拡張された手法が最も良い成績を残している [8]．

さらに，機械学習手法として決定リストを用いる理由の1つに，決定リストは優先順位を与えられたif-then形式の規則の並びであるため，学習結果や適用結果が容易に考察できるということがある．今回参加した辞書タスクでは，訓練データと評価データの両方に国際十進分類法（Universal Decimal Classification）[9] による文書分類コード（以下，UDCコード）が付与されている．また，語義の定義を行っている岩波国語辞典の語釈文を判定に利用することも考えられる．したがって，考察が容易にできるという決定リストの利点を生かして，これらの情報を用いた判定がどの程度有効であるかを実験的に検証した．

本論文の構成を以下に示す．2節では，決定リスト作成の際に証拠としてどのような情報を用いたかなど，語義曖昧性解消を行うシステムの作成手法について述べる．3節では，作成したシステムを辞書タスクの評価データに適用して実験を行ない，その結果を示す．最後に，4節で今後の課題について述べる．

2. 語義曖昧性解消システム

本節では，決定リストを用いて語義曖昧性解消を行うシステムの作成手法について述べる．日本語の辞書タスクは，語義曖昧性解消を行う対象となる単語があらかじめ設定されているlexical-sampleタスクである(注1)．今回のコンテストでは，100種類の単語（名詞50種類，動詞50種類）について，各単語100事例ずつ，合計10,000個の事例が評価データとなっている．したがって，各単語ごとにひとつずつ，合計100個の決定リストを訓練データから学習した．以降では，決定リストの学習アルゴリズムについて述べる．

2.1 証拠の生成

ここでいう証拠とは，語義曖昧性解消に用いる文脈情報のことである．決定リストに限らず多くの機械学習手法では，クラス分類にどのような証拠を用いるかが重要になる．本節では，我々が用いた証拠について述べる．

辞書タスクにおける訓練データは，新情報処理開発機構（RWCP: Real World Computing Partnership）テキスト・サブ・ワーキンググループによって作成された語義タグ付きテキストデータベースである [5]．これは，毎日新聞の新聞記事3,000個からなり，各記事は自動的に形態素解析された後，人手によって修正されている．また，それぞれの記事には文書分類を表すUDCコードが付与されている．さらに，構文解析器KNPパーザを用いて各文章に構文情報を追加し，最終的には，形態素情報，構文情報，文書分類情報の3つから証拠を生成することにした．

（1） 形態素情報

形態素情報から生成する証拠は，さらに，Adjacent，Pair，Windowの3つに分類される．括弧内の記号は，それぞれの証拠を略記したものである．また，証拠に活用語が含まれる場合には，与えられている形態素情報から活用語の基本形に直してから証拠を生成することとする．これは構文情報から証拠を生成する場合も同様である．

- Adjacent
  対象単語の直前に現れる単語（W_{-1}）．
  対象単語の直後に現れる単語（W_{+1}）．
  対象単語の直前に現れる単語の品詞（P_{-1}）．
  対象単語の直後に現れる単語の品詞（P_{+1}）．
- Pair
  対象単語の左の単語対（$W_{-2,-1}$）．
  対象単語を挟む単語対（$W_{-1,+1}$）．
  対象単語の右の単語対（$W_{+1,+2}$）．
  対象単語の左の品詞対（$P_{-2,-1}$）．

(注1)：全ての単語を対象とするall-wordsタスクが用意されていた言語もある．

対象単語を挟む品詞対（$P_{-1,+1}$）.
対象単語の右の品詞対（$P_{+1,+2}$）.

- Window
  対象単語から $\pm k$ 単語以内に現れる単語.
  今回の実験では $k = 4$ とした（Win）.

例えば,「麦$_1$: から$_{419}$: ビール$_1$: を$_{419}$: 作る$_{271}$: ◦$_{468}$」(注2)という文で対象単語が「作る」の場合には，以下の証拠が生成される.

（W_{-1} を）,（W_{+1} ◦）,（P_{-1} 419）,（P_{+1} 468）,（$W_{-2,-1}$ ビール を）,（$W_{-1,+1}$ を ◦）,（$P_{-2,-1}$ 1 419）,（$P_{-1,+1}$ 419 468）,（Win 麦）,（Win から）,（Win ビール）,（Win を）,（Win ◦）

（2） 構文情報

構文情報から生成する証拠は，対象単語の品詞によって区別する.

まず，対象単語が名詞の場合には，その名詞が係っている動詞を証拠として生成する．例えば,「後ろ:に:手:が:回る:◦」という文で対象単語が「手」の場合には,（Syn_N 回る）という証拠が生成される.

対象単語が動詞の場合には，その動詞のヲ格の格要素を証拠として生成する．格要素が複合名詞ならば，生成する証拠の種類が増えるのを抑えるために，複合名詞全体を参照せず，複合名詞の一番最後に現れる名詞のみを参照する．例えば,「製造:許可:を:与える:◦」という文で対象単語が「与える」の場合には,（Syn_V 許可）という証拠が生成される.

（3） 文書分類情報

UDC コードは 10 進数字と若干の記号によって表示される．その数字的表示を標数と呼ぶ．標数には，主標数，固有補助標数，共通補助標数の区別があり，それらを組み合わせて分類を行う．今回の実験では，簡単のため主標数のみを取り出して使用した.

2.2 岩波国語辞典の利用

今回配布された訓練データは，1 つの対象単語あたりの事例数が最小 71 個から最大 911 個と少ない．そこで，少しでも訓練データを増やすために，岩波国語辞典の語釈文中に含まれている例文を使用した．語釈文中の例文には，その語義における代表的な使用例が挙げられているため，語義曖昧性解消に対して有効に働くと考えられる.

(注2)：":" は単語区切り，添字はその単語の品詞コードを表わしている.

語釈文には，訓練データと同様に形態素解析された情報が付与されているので，KNP パーザを用いて構文解析した後，2.1 節で述べた形態素情報と構文情報から得られる証拠を生成する．UDC コードは付与されていないので，文書分類情報から証拠を生成することはできない.

2.3 決定リスト作成

2.1 節，2.2 節のようにして訓練データと語釈文中の例文から生成された証拠 E_i から，式 1 の規則 r_i を作成する.

$$r_i: \quad E_i \quad \rightarrow \quad s_i \tag{1}$$

ここで s_i は，証拠 E_i が成立するときの語義を表す．また，式 2 で表されるデフォルト規則を導入する.

$$true \quad \rightarrow \quad s_{mfs} \tag{2}$$

これは対象単語に対して無条件で適用される規則である．s_{mfs} には，対象単語ごとに最も頻度の高かった語義を与える.

こうしてできた規則の適用順序を決め，決定リストを作成する．規則の適用順序は規則の尤度に従う．規則 r_i の尤度 $L(r_i)$ は式 3 で定義される.

$$L(r_i) = log\frac{P(s_i|E_i)}{P(\overline{s_i}|E_i)} \tag{3}$$

ここで $P(s_i|E_i)$ は，証拠 E_i が成立するときに語義が s_i になる確率であり，$P(\overline{s_i}|E_i)$ はその排反事象である．これは式 4 で推定する.

$$P(s_i|E_i) = \frac{O(E_i, s_i) + \alpha}{O(E_i) + \alpha} \tag{4}$$

ここで $O(E_i)$ は証拠 E_i が成立する文脈の出現頻度であり，$O(E_i, s_i)$ は証拠 E_i が成立し，かつ対象単語の語義が s_i である文脈の出現頻度である．また，α はスムージングのためのパラメータである．今回の実験では，$\alpha = 0.5$ として使用した.

このように作成された決定リストは式 2 のデフォルト規則を含んでいる．デフォルト規則は対象単語に対して無条件で適用される規則であるため，デフォルト規則より優先順位の低い規則が適用されることはない.

3. 評価実験

3.1 学　　習

評価実験では，語釈文中の例文を訓練データとして使用することがどの程度有効であるかを調べるために，訓練データのみから学習した決定リストと，訓

練データと語釈文中の例文から学習した決定リストの 2 つを作成した．以降では，前者を DL，後者を DL+と表記する．

3.2 評価データ訂正

前述したように，訓練データは自動的に形態素解析された後，人手による修正が加えられている．それに対して，評価データは自動的に形態素解析されただけである．これでは，学習した決定リストを適用する際に，わかち書きや品詞の違いから間違った文脈情報が得られてしまう可能性が高い．そこで，人手が加えられる前の訓練データと人手が加えられた後の訓練データを比較して，わかち書きと品詞を訂正する規則を取り出し，評価データに適用することを考えた．

まず，わかち書きの訂正規則は，人手が加えられる前後の訓練データを比較してわかち書きが異なる部分から取り出した．このとき，前後の文脈や修正前のわかち書きにおける各単語の品詞は利用しない．取り出された訂正規則の例を挙げる．

$$\text{と}:\text{し}:\text{て} \quad \rightarrow \quad \text{として}_{421} \tag{5}$$

$$\text{容疑者} \quad \rightarrow \quad \text{容疑}_{1}:\text{者}_{24} \tag{6}$$

取り出された訂正規則は，出現頻度により優先順位を与えて上位の訂正規則から順に適用できるものすべてを適用した．ただし，頻度が 2 回以下の訂正規則は除いている．

次に，品詞の訂正規則は，人手が加えられる前の訓練データのわかち書きを変更し，それと人手が加えられた後の訓練データを比較して品詞が異なる部分から取り出した．ここでは，直前の単語の品詞と直後の単語の品詞を文脈として利用した．取り出された訂正規則は，わかち書きの訂正規則と同様に適用した．ただし，頻度が 60 回以下の訂正規則は除いている．

3.3 実験結果と考察

表 1 実験結果

正解率	BL	W	DL	DL+
評価データ訂正前	72.6%	77.1%	76.77%	77.79%
訂正後	72.6%	77.0%	76.78%	77.91%
規則適用率	BL	W	DL	DL+
評価データ訂正前	0%	93.09%	97.97%	97.99%
訂正後	0%	93.93%	98.87%	98.87%

表 1 は，学習した 2 つの決定リストをそれぞれ訂正前後の評価データに適用したときの正解率と規則適用率である．正解率は語義を正しく判定できた割合，規則適用率は，正誤に関わらずデフォルト規則以外の規則によって語義を判定できた割合である．BL（ベースライン）は，対象単語ごとに最も頻度の高い語義を選択する手法を表す．W は 7 月に開催されたワークショップに提出したときの手法を表す．いずれの手法の正解率もベースラインを 4，5%ほど上回った．DL と DL+の正解率を比較すると，DL+のほうが約 1%良くなっている．これは，語釈文中の例文が適当な訓練データとなりうることを示している．一方，訂正前の評価データに適用した結果と，訂正後の評価データに適用した結果を比較しても，ほとんど差は見られなかった．DL+がワークショップ時の手法と比べて良くなっているのは，ワークショップ時には考慮されていなかった構文情報から得られる証拠を追加したからであると考えられる．その他に文書分類情報から得られる証拠のバグフィックスも行ったが，後述するように文書分類情報に基づいた規則が適用される回数は非常に少ない．したがって，構文情報から得られる証拠が有効に働いたと考えるほうが自然である．

表 2 規則のタイプ別正解率

規則のタイプ	適用数	正解率
単語 *Adjacent*	1030	83.59%
品詞 *Adjacent*	971	84.65%
単語 *Pair*	796	74.25%
品詞 *Pair*	1515	78.22%
Window	3507	77.99%
Syn_N	707	76.66%
Syn_V	817	78.09%
UDC	544	63.42%
デフォルト規則	113	63.72%

表 2 は，訂正後の評価データに対して DL+を適用した際に，実際に適用された規則ののべ数とその正解率を規則のタイプ別に示したものである．直前，直後の単語もしくは品詞のみを用いた規則が最も良い正解率を示している．文書分類情報から得られる証拠は適用数が少なく，その正解率も低い．

最後に，訂正後の評価データに対して DL+を適用した際の単語別正解率を対象単語の品詞と難易度ごとに分けて図 1 から図 6 に示す．各単語とも左側の棒グラフがベースラインの正解率を，右側の棒グラフが決定リストを適用したときの正解率を表している．「目」「地方」「近く」など正解率が 10%以上改善されている単語が 25 個あるが，一方では，「情報」「同日」「今」などベースラインよりも正解率が悪くなっている単語も 12 個ある．

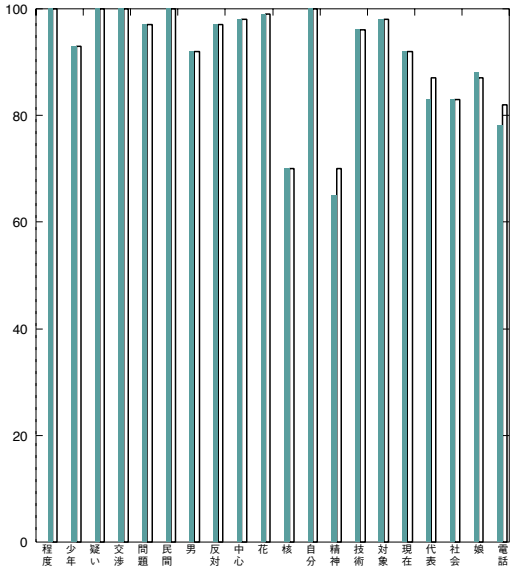

図 1　名　　詞:易

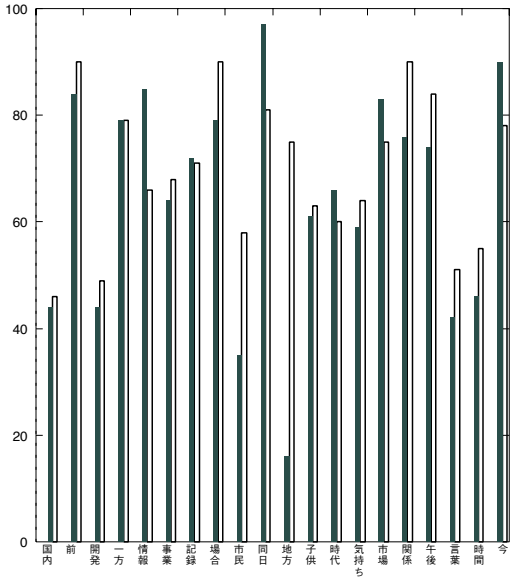

図 2　名　　詞:中

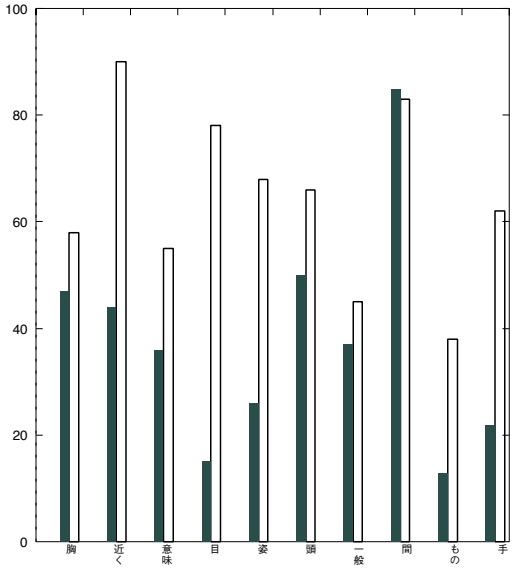

図 3　名　　詞:難

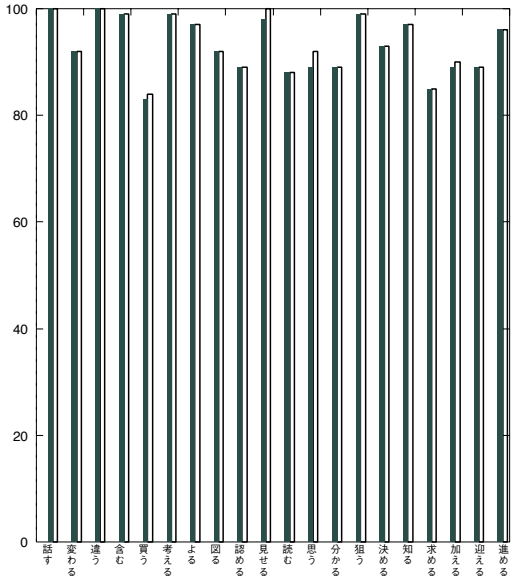

図 4　動　　詞:易

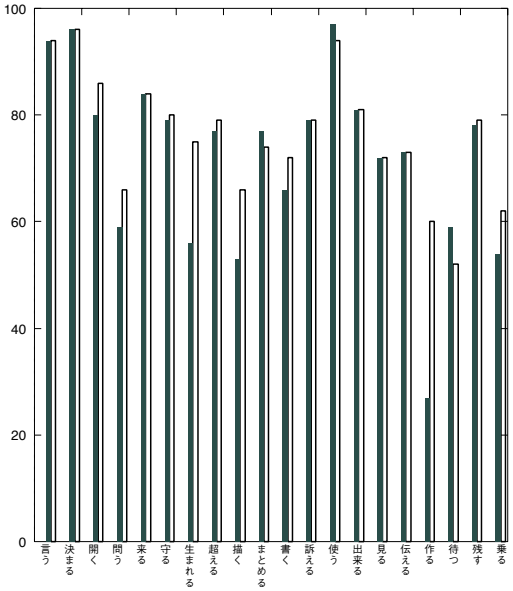

図 5　動　　詞:中

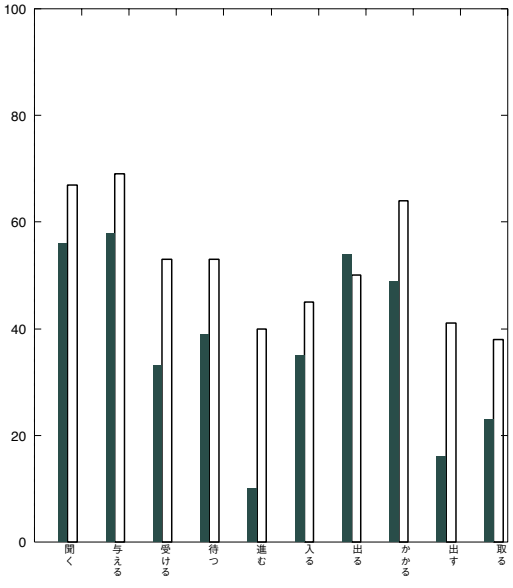

図 6　動　　詞:難

4. おわりに

訓練データの不足を補うために語釈文中の例文を利用することによって，語義曖昧性解消の正解率が上昇することを示した．それでも最終的な結果では，1 回しか出現していない証拠に基づいた規則が適用される場合が数多くみられた．したがって，今後は，意味クラスを用いて証拠に含まれる単語を抽象化することを考えている．しかし，白井らは証拠に意味クラスを導入するナイーブな手法では結果が悪くなるため，導入の際に語義曖昧性解消を行う必要があると述べている．そこで，訓練データに付与されている語義を利用しつつ，決定リストの作成に意味クラスを導入する方法を検討したい．

また，今回は語釈文中の例文のみを利用するに留まったが，岩波国語辞典の語義の定義文を利用することも考えている．

文　献

[1] Ronald L. Rivest. Learning decision lists. *Machine Learning*, Vol. 2, pp. 229–246, 1987.

[2] 颯々野学, 宇津呂武仁. 統計的日本語固有表現抽出における固有表現まとめ上げ手法とその評価. 情報処理学会自然言語処理研究会, Vol. 2000, No. 86, pp. 1–8, 2000.

[3] 新納浩幸. 決定リストを弱学習器としたアダブーストによる日本語単語分割. 自然言語処理, Vol. 8, No. 2, pp. 3–18, 2001.

[4] 白井清昭, 橋本泰一, 西舘耕介, 徳永健伸, 田中穂積. 決定リストを用いた形容詞の修飾先の決定. 言語処理学会 第 7 回年次大会 発表論文集, pp. 253–256, 2001.

[5] 白井清昭, 柏野和佳子, 橋本三奈子, 徳永健伸, 有田英一, 井佐原均, 荻野紫穂, 小船隆一, 高橋裕信, 長尾確, 橋田浩一, 村田真樹. 岩波国語辞典を利用した語義タグ付きテキストデータベースの作成. 情報処理学会自然言語処理研究会, Vol. 2001, No. 9, pp. 117–122, 2001.

[6] David Yarowsky. DECISION LISTS FOR LEXICAL AMBIGUITY RESOLUTION: Application to Accent Restoration in Spanish and French. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 88–95, 1994.

[7] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of theAnnual Meeting of the Association for Computational Linguistics*, pp. 189–196, 1995.

[8] David Yarowsky. Hierarchical Decision Lists for Word Sense Disambiguation. *Computers and the Humanities*, Vol. 34, No. 1–2, pp. 179–186, 2000.

[9] 情報科学技術協会. 国際十進分類法. 丸善株式会社, 第 3 版, 1994.