IPAL 辞書の自動的ハイパーテキスト化

梁 慶昇, 奥村 学
北陸先端科学技術大学院大学 情報科学研究科
email:{ryou,oku}@jaist.ac.jp

1 はじめに

自然言語処理では膨大な情報を持つ辞書は不可欠である。自然言語処理に用いることを目的として通産省の外郭団体である情報処理振興事業協会 (IPA) は計算機用辞書 IPAL 辞書 [2][3][4] を作成している。この辞書の詳細な情報は研究者にとって役立つが、3つの辞書に分かれているため、参照しにくい。また辞書間に名詞の意味素性における不統一性が存在している。

辞書の互いに参照しにくい問題を解決できる手法としてハイパーテキストがある。ハイパーテキスト [1] とはテキストを意味上まとまりのあるブロックに分割し、それぞれの間にリンクにより関係づけたネットワーク構造で情報を管理する技術である。

そこで本研究では IPAL 辞書をハイパーテキスト化する。ハイパーテキスト化の過程で作成した単語のネットワーク構造により、意味素性の不統一性を解消する。これによって、IPAL 辞書の利用者に利用しやすい枠組を提供する。

2 IPAL 辞書のハイパーテキスト化

システムのイメージを図 1に示す。検索したい単語を入力すると、入力した単語に対して、単語の持つ情報の構造を表示し、参照したい項目をクリックすると、項目の持つ情報を表示する。さらに表示された情報内のノードをクリックすると素早く単語の情報を参照出来る。しかし一般に単語には複数の語義があるため、本研究はこれらの語義の中から正しい語義を選び出して、リンクを張り、名詞の意味素性における不統一性を解消する。

そのためには、見出し語の情報内の“ノードの決定”

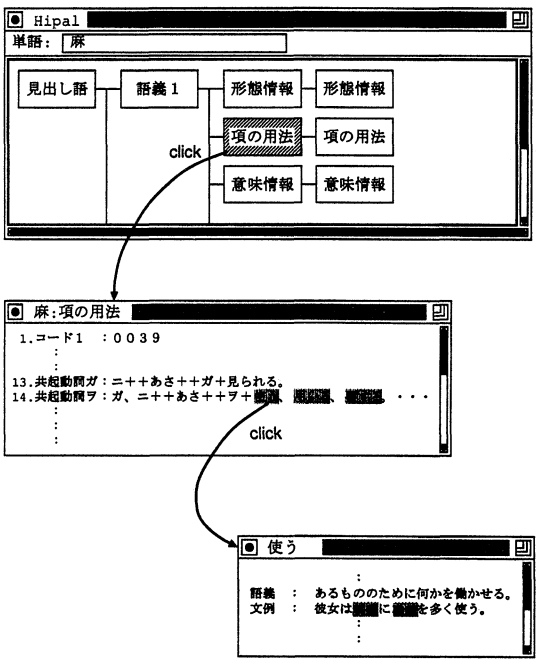


図 1: システムのイメージ図

と“リンクによって張られた単語の語義決定”という2つの処理が必要である。

2.1 ノードの決定

見出し語の情報より関連性（複合語、共起単語、類似語など）を持つ単語を自動的に発見し、ノードとしてリンクを張る。具体的な手順は次のようになる。

- 辞書の見出し語ごとの情報を項目単位に分割する。
- 日本語解析の必要な箇所（例文、名詞句など）は、日本語形態素解析システム JUMAN[5] を用いて、

単語列への分割を行なう。

- 見出し語と関連性のある単語をノードとする。
- ノードにあたる単語から辞書の見出し語にリンクを張る。

2.2 語義決定

ノードから辞書の見出し語にリンクを張る際、張られた単語の正しい語義を決定する。本研究はノードが共起単語の場合だけ語義決定を行なう。また名詞辞書の共起動詞、共起形容詞、共起名詞と動詞、形容詞辞書の共起名詞の５種類の共起単語の語義決定の方法を“名詞辞書から他辞書”と“他辞書から名詞”の２つの場合に分け、それぞれ４つ、５つの手法を検討した。それぞれの手法は以下の通りである。

[“名詞辞書から他辞書”の場合]

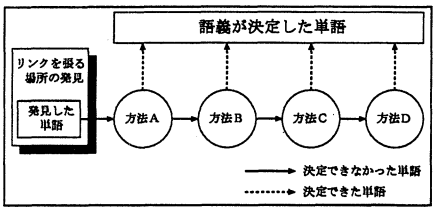
1. 名詞句による語義決定
2. 意味素性の対応表（IPAが提供した辞書間の名詞に振った意味素性における違いを記述した表）に基づく語義決定
3. 意味素性の完全一致による語義決定
4. 文型による語義決定

[“他辞書から名詞辞書”の場合]

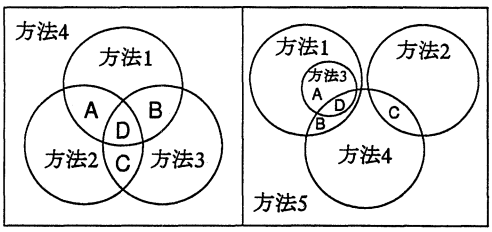
1. 動詞の意味分類番号（角川類語辞典[6]）による語義決定（３桁以上）
2. 動詞の意味分類番号による語義決定（２桁）
3. リンク・テーブル（単語のネットワーク構造を保存する場所）の完全マッチによる語義決定
4. リンク・テーブルの名詞の候補による語義決定
5. 意味素性の対応表を利用して、語義決定を行なう

2.3 適用順序の決定

１つの方法では、すべての共起単語の語義を決定できない。そこで、これらの方法を組み合わせて語義を決定する。本研究では語義を決定するときに、図のように方法Aで語義が決定できるなら、方法B～Dを適用しないため、方法を実行する順序を決定する必要がある。ここで、名詞辞書から他辞書の方法４と他辞書から名詞辞書の方法５は複数の語義から一意に決定できない場合があるため、組合せの最後にする。



他の方法を実行する順番を決定するにはさきほども説明したように、それぞれの方法は全共起単語の内、カバーする所に偏りがある。その内、２個以上の方法が決定できる部分を共通部分と呼ぶ。また１個の方法しか決定できない部分を独立部分と呼ぶ。独立部分は全体の決定精度に影響しないが、共通部分は精度の高い方法を先に実行することによって、全体の精度を上げることができるので、共通部分を評価して、順番を決定する。それぞれの場合の共通部分は図のようになる。



名詞辞書から他辞書　　他辞書から名詞辞書

共通部分における評価（名詞辞書から他辞書）				
共通部分	評価数	不正解の数		
		方法１	方法２	方法３
A	876	68	53	
B	1379	107		260
C	1116		50	50
D	1052	90	28	28

共通部分における評価（他辞書から名詞辞書）					
共通部分	評価数	不正解の数			
		方法１	方法２	方法３	方法４
A	78	1		1	
B	29	3			4
C	35		7		5
D	126	6		6	19

評価した結果を見ると順番２→１→３→４、１→３→４→２→５が一番良い精度を得られると期待できる。

3 実験

実際に提案した方法を組み合わせて全共起単語の語義を決定して評価した結果、表1のように、やはり２→１→３→４と１→３→４→２→５が一番良い精度を得た。

方法の組合せの評価 “名詞辞書から他辞書”								
組合せ	評価数	方法1	方法2	方法3	方法4	その他	合計	正解率
1234	1018	53	18	13	12	12	108	89.4%
1324	1018	53	18	13	12	12	108	89.4%
2134	1018	42	25	13	12	12	104	89.8%
2314	1018	26	25	44	12	12	119	88.3%
3124	1018	31	18	48	12	12	121	88.1%
3214	1018	26	21	48	12	12	119	88.3%

方法の組合せの評価 “他辞書から名詞辞書”									
組合せ	評価数	方法1	方法2	方法3	方法4	方法5	その他	合計	正解率
13245	790	13	8	0	0	0	5	26	96.7%
13425	790	13	6	0	0	0	5	24	96.9%
31245	790	10	8	3	0	0	5	26	96.7%
31425	790	10	6	3	0	0	5	24	96.9%
34125	790	9	6	3	1	0	5	24	96.9%
41325	790	12	6	0	1	0	5	24	96.9%

表 1: 方法の組合せの評価

4 考察

上の方法を使って、語義を決定する場合に正しい語義を決定できない原因は、以下のようなものがある。ここで名詞辞書から他辞書と他辞書から名詞辞書の２つの場合について述べる。

[“名詞辞書から他辞書” の場合]

1. 具体名詞を抽象名詞として使ったため、失敗する。（方法１）
2. 意味素性の対応表以外の対応がある。（方法２）
3. 正しい語義の共起名詞の意味素性は名詞の見出し語の意味素性の下位、上位概念、または違う概念になっている。（方法３）
4. 文型が違う。（方法４）
5. 正しい語義がない。
6. 格形式による失敗
7. 動詞の漢字の違いによる失敗

この７つの正しい語義が決定できない原因を見ると、原因１～４はそれぞれ方法１～４に当たる。原因５～７はIPAL辞書の問題で、もし作成者がこの問題を改善できるなら、もっと良い精度が期待できる。

[“他辞書から名詞辞書” の場合]

1. 動詞単語の語義について考慮していないため、失敗する（方法１、２）

   方法１、２は意味分類番号を使って意味的に近い単語を集めるが、単語は複数の語義を持つのが一般的であり、語義決定を行なうときに、単語の表記のみを利用するから、誤った語義を選んでしまう。

2. もともとリンク・テーブルが間違っている（方法３）

   方法３は名詞辞書から他辞書にリンクを張った結果を利用するが、リンク・テーブルがそもそも間違っている場合がある。

3. 名詞の候補の語義について考慮していないために失敗する（方法４）

4. 意味素性の対応表以外の対応があり、決定できない場合がある（方法５）

5. 名詞句を単語列に分割することによる失敗

6. 漢字の読みの違いによる失敗

この６つの原因のうち原因１から原因４は方法１～方法５に当たる。原因５はJUMANが名詞句を単語列に分割するときに起きた失敗、６は漢字が同じだが、読みが違う際に起こる失敗である。

5 ブラウズシステム

“ノードの決定”と“語義決定”の結果を用いてワークステーション上で利用できるハイパーテキスト・ブラウズシステムを作成した。これらはTcl/TkとC言語で記述されている。システムは関連する単語の間にリンクを張ったハイパーテキスト構造によって辞書を管理する。主な機能は以下の通りである。

- IPAL辞書は見出し語ごとの情報が極めて詳細である。しかし、利用者ごとの要求が様々であるため、参照したい情報も違ってくる。そこで、システムの機能として利用者の参照したい情報だけを表示するようにする
- 見出し語として用意されていない単語を検索した場合、その単語が辞書内で使用されていれば、その単語がどの見出し語でどのように使われているかを参照できる。これによって作成者が新たに見出し語を追加する場合に参考となる情報を提供できる
- ユーザが設定したカスタマイズ情報を元に、辞書の見出し語ごとの情報をファイルに保存することができる。この機能は辞書の特定な情報しか利用しない人に有効である

6 結論

本稿ではIPAL辞書のハイパーテキスト化に関する研究を行なった。ハイパーテキストはIPAL辞書のような大規模な辞書を効率よく管理出来る。またリンクによって張られる単語の語義を複数の語義の中から正しい語義だけを選び出して張ることにした。そして語義の決定結果は８９％以上の精度が得られた。これにより、辞書間に存在する名詞の意味素性の不統一性を解決できた。さらにブラウズ・システムの様々な機能によって、作成者に対してもより良い環境を提供した。これらの結果から見ると本研究の目的である作成者や利用者に参照・利用しやすい枠組を提供できた。

今後の課題としては次の３つがある。

1. 今回の研究でリンクの張る先の単語の複数の語義から正しい語義を決定するとき、複数の語義を決定してしまう場合がある。このような場合を単独に決定する手法を開発する
2. リンク・テーブルはノードとノード間のリンク関係を記述するものである。IPAL辞書の見出し語ごとの情報が詳細であるため、リンク・テーブルの容量もかなり大きい（４M bytes）。また現在のIPAL名詞辞書に載っている見出し語の数が８０５語であり、今後辞書のバージョンアップが行なわれると、リンク・テーブルの容量も増える。そこで、リンク・テーブルの圧縮が必要となってくる
3. インターフェイスの評価は時間を要する事であるため、今回の研究では十分行なうことができなかった。作成したインターフェイスを実際に作成者や利用者が使って、作成者や利用者の声を聞いてから、インターフェイスを改良する

謝辞

貴重な御意見、御討論をいただきましたIPAの方々に心から御礼を申し上げます。

参考文献

[1] 黒橋 禎夫、長尾 真、佐藤 理史、村上 雅彦, “専門用語辞典の自動的ハイパーテキスト化の方法”, 人工知能学会誌 Vol.7 No.2,page 336-345, 1991.

[2] 情報処理振興事業協会技術センター, “計算機用日本語基本動詞辞書IPAL解説”, 1987.

[3] 情報処理振興事業協会技術センター, “計算機用日本語基本形容詞辞書IPAL解説”, 1990.

[4] 情報処理振興事業協会技術センター, “計算機用日本語基本名詞辞書IPAL解説”, 1994.

[5] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真, “日本語形態素解析システムJUMAN使用説明書 version 2.0”, 京都大学工学部長尾研究室, 奈良先端科学技術大学院大学松本研究室, 1994.

[6] 大野 晋, 浜西 正人, “角川類語新辞典”, 角川書店, 1981.