EDR電子化辞書からの意味空間の構築とWeb情報による再構築

Construction of the semantic network from EDR Electronic Dictionary, and Reconstruction using Web information

墨岡 沖* 中山 尭†

概 要

人間は言葉によって互いに意志疎通を行なっている．これは人間の脳の中で言葉と概念を無意識に繋げる事で行なわれる．この意味空間は外界との会話などの刺激によってたえず変化している．

本研究ではEDR電子化辞書 [1] を用いて計算機上に意味空間を作成する．次に外界からの入力を情報の豊富なWeb資源から自動的に行なう．これにより意味空間がどのように変化し再構成されるのかを観察し考察する．意味空間に新語 (新概念) を与えた時，その語がシソーラス上のどこに配置されるかを見る事でWeb情報の活用性，信用性などを考察する．

1. はじめに

人間は無意識にものを考え，言葉によって他の人間とコミュニケーションを取っている．脳の中に意味空間を持ち，絶えず新しい情報を取り入れて再構成している．意味空間と語彙を結びつけることで発話も可能となる．しかし，これを計算機上で行なうことは難しい．

本研究では既存の辞書から意味空間の初期構造を作成する．これに，Webから自動検索した入力文を与え意味空間を再構成し，学習するシステムを構築する．これによりWeb情報の活用性，信用性なども考察する．

*北陸先端科学技術大学院大学情報科学研究科情報システム学専攻 修士課程

†神奈川大学理学研究科教授 理学博士

2. 研究背景

今やWebは情報発信の手段として多く利用されている．個人レベルの小さなWebサイトから大手企業のWebサイトまで様々な情報が乱雑に広がっている．そのWeb情報資源の量は膨大であるが，間違っている情報も多数存在する．

本研究では意味空間をEDR電子化辞書から抽出し，Web情報を利用して意味空間の再構成を試みる．これによりWeb資源の活用性，信用性などがどの程度有効であるか考える．

3. 先行研究

Web資源から特定の情報を抽出する研究は近年盛んにされている．特定の語の情報をWebから収集したり，Web上の文章の要約などが一例として挙げられる．また,Webにおけるオントロジー記述言語としてOWL(Web Ontology Language)[2] が開発され，英語シソーラスであるWordNet.OWL[3] が公開されている．

しかしEDR電子化辞書とWebの組合せをターゲットとしたものは少なく，EDR電子化辞書から意味空間を構築しWebを用いて意味空間を再構築する

システムは存在しない.

4. システム概要

当システムは大きく 2 つの作業に分けられる．一点目は “意味空間の構築”．二点目は “Web 情報による再構築” である．section5 で一点目を，section6 で二点目を詳しく説明する．下記に 2 つの作業の概要を述べる．また，図 1 はシステム構成図を示している．

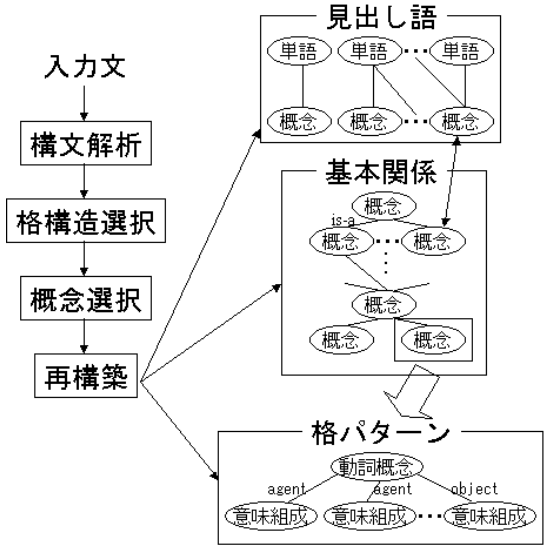

図 1: システム構成

4.1. 意味空間の構築

まず,EDR 電子化辞書データベースを SQL 上に構築する．SQL サーバは PostgreSQL を用いる．SQL から抽出し新しい 3 つのテーブル “見出し語”,“基本関係”,“格パターン” を作成する．この 3 つのテーブルを意味空間と呼ぶ．これは図 1 では右側に示している．

4.2. 意味空間の再構築

Web からの入力文により，意味空間が再構築される．構文解析，意味解析を経て意味空間に入力される．本論文では動詞のみをターゲットとする．入力文の動詞と共起する語は，その動詞の意味素性に関係があると考える．入力語が意味素性に含まれればそのリンク強度を強くし，含まれなければ新たにリンクを作る．多くの入力があれば適切なリンクのみ強くなると考えられる．この操作は図 1 の左側に示されている．

5. EDR 電子化辞書からの意味空間の構築

意味空間の入力として EDR 電子化辞書を用いる．EDR 電子化辞書は日本語の電子辞書としては最も大きいものであり，本研究を遂行するには適切であると考えられる．EDR 電子化辞書は各辞書ごとにテキストファイルになっており，一行を一項目という形式で書かれている．まず，EDR 電子化辞書を section5.1 のように第三正規形まで正規化しデータベースを PostgreSQL 上に置いた．正規化は一般的に冗長なデータを減らす為に行なわれる．正規化により一意性制約や参照制約などにより検索が正しく行なわれる．反面，検索時間は増加する傾向がある．PostgreSQL はフリーの SQL サーバであるが，充分実用に耐え得る物とされている．構築は EDR 電子化辞書のテキストファイルからスクリプトを用いて各行を読み込み，SQL サーバに入力される．

SQL 上に構築された EDR 電子化辞書から意味空間を作成する．意味空間は “見出し語”，“基本関係”，“各パターン” の 3 つのテーブルで定義する．

5.1. 第三正規形

EDR 電子化辞書は全く正規化されていない為，正規化が必要であると考えられる．一般的に正規化することで冗長なデータを減らすと共に制約により正

表 1: 第三正規形

概念体系辞書	
概念体系辞書	レコード番号 上位概念識別子 下位概念識別子 管理情報
概念見出し辞書	
概念見出し辞書	レコード番号 概念識別子 英語概念見出し 日本語概念見出し 英語概念説明 日本語概念説明 管理情報
概念記述辞書	
概念記述辞書	レコード番号 記述区分 概念識別子１ 関係子 概念識別子２ 真偽値 管理情報
英語共起辞書	
英語共起辞書	レコード番号 単語表記１ 共起関係子 単語表記２ 受け側要素 関係要素 係り側要素 受け側概念要素 概念関係子 係り側概念要素 頻度例文 管理情報
英語共起句構成要素情報	レコード番号 要素番号 形態素 原形 品詞 慣用句フラグ 概念識別子 補足付き概念説明
英語コーパス	
英語コーパス	レコード番号 テキスト番号 出典情報 文 形態素情報 構文情報 意味情報 管理情報
英語コーパス構成要素情報	レコード番号 構成要素番号 表記 表記原形 品詞 概念選択
英日対訳辞書	
英日対訳辞書	レコード番号 単語見出し 品詞 概念識別子 対訳情報 管理情報
英語単語辞書	
英語単語辞書	レコード番号 単語見出し 不変化部_連接属性対 音節区切り 発音 品詞 構文木 語形・語形変化情報 文法属性 文型情報 機能・位置 機能語情報 概念識別子 用法 頻度 管理情報
日本語共起辞書	
日本語共起辞書	レコード番号 単語表記１ 共起関係子 単語表記２ 受け側要素 関係要素 係り側要素 受け側概念要素 概念関係子 係り側概念要素 頻度例文 管理情報
日本語共起句構成要素情報	レコード番号 要素番号 形態素 かな表記 品詞 慣用句フラグ 概念識別子 補足付き概念説明
日本語動詞共起パターン副辞書	
日本語動詞共起パターン副辞書	レコード番号 文パターン見出し 例文 管理情報
日本語動詞共起パターン構成要素	レコード番号 要素番号 表記 文法情報 概念関係子 概念識別子
日本語動詞共起構文情報構成要素	レコード番号 概念関係子 格助詞
日本語動詞共起意味情報構成要素	レコード番号 概念関係子 概念識別子 概念説明
日本語コーパス	
日本語コーパス	レコード番号 テキスト番号 出典情報 文 形態素情報 構文情報 意味情報 管理情報
日本語コーパス構成要素情報	レコード番号 構成要素番号 表記 かな表記 品詞 概念選択
日英対訳辞書	
日英対訳辞書	レコード番号 単語見出し 品詞 概念識別子 訳語情報 管理情報
日本語単語辞書	
日本語単語辞書	レコード番号 単語見出し 不変化部_連接属性対 かな表記 発音 品詞 構文木 活用情報 表層格情報 相情報 機能語情報 概念識別子 用法 頻度 管理情報

しい検索を行なう事が可能となる．データベースの正規化は第五正規形まであるとされているが，実用上は第三正規形までで十分と言われている．本システムでは EDR 電子化辞書を第三正規形まで正規化を行なった．具体的には，まず繰り返し項目の削除を行ない第一正規形にする．次に部分関数従属の排除を行ない第二正規形にする．そして推移関数従属の排除を行ない第三正規形にする．これは決まった手法で行なうが計算機で自動的に行なう事は難しく，人間の手により行わざるを得ない．ここで得られた第三正規形のテーブルを表 1 にまとめた．

5.2. 意味空間の構築

EDR 電子化辞書の概念見出し辞書，概念体系辞書，日本語動詞共起パターン副辞書から抽出し，“見出し語”，“基本関係”，“格パターン” を作成する．見出し語 40 万，基本関係 40 万，格パターン 5 万の意味空間を作成する．意味空間は下記の様な様式である．

5.2.1. “見出し語”

- 見出し語 (読み 書き 概念)

単語見出し (読み，書き) と概念とのリンクを記述する．EDR 電子化辞書の日本語単語辞書から構築する．

5.2.2. “基本関係”

- 基本関係 (上位概念 下位概念 重み)

概念の上位下位の関係を記述する．EDR 電子化辞書の概念体系辞書から構築する．ここで言う概念の上位下位とは包含関係 (is-a) を指す．リンクには重みが設定されている．これは 0 から 1 までであり，1 に近いほどリンクは正しいとされる．重みがしきい値以上の時正式なリンクとして認められる．当論文の実験では有効となるしきい値を算定していない為しきい値は考えない．EDR 電子化辞書のデータは絶対に正しいという前提があるとし，基本概念の初期の重みは最大値である “1” とする．

5.2.3. “格パターン”

- 格パターン (概念 関係子 意味素性)

動詞概念の関係子，意味組成を記述する．EDR 電子化辞書の概念記述辞書から構築する．動詞概念には agent goal object place など 9 種類の関係子を与える．この関係子は例えば agent は “動作を引き起こす主体” などである．意味組成はそれぞれについてなりうる概念である．可能性のある全ての概念を記述するのではなく “基本関係” を参照し is-a の最上位のみを記述する．

6. Web 情報による再構築

6.1. 入力文の取得

ロボット型検索エンジンの一つである google を利用する．Web の検索エンジンの種類はロボット型と登録型の二種類に大別される．再構築の際使う入力文は多岐にわたる物が適切だと思い前者であるロボット型検索エンジンを用いる．google に対して特定の “語” を検索させ “語” を含む文を入力文として利用する．検索結果のページのアンカー先にはその “語” に関連するサイトがあるという予測ができるからである．初期 URL(google の提示した URL) から最大 5 ホップ (リンクを 5 つ辿った) 先までの URL を自動検索させ，入力文を集める．

6.2. 意味空間への挿入

入力文は大多数になる為，JUMAN[4] で形態素解析，KNP[5] を用いて構文解析を行ない自動化する．意味解析は辞書を利用することにより半自動的に行なう．一つしか概念のない語は自明であるが，複数の概念を持つものは手動で判別する．格構造選択も同様の手法を用いる．最終的に格パターンの “概念”，

“関係子”，“意味組成” の 3 つ組が入力される．このとき意味組成は基本関係に反映され重みの再構成が行なわれる．

7. 実験

新語として “紙” を追加した．“紙” という語を選ぶにあたっては特に理由はないが，十分一般的な語と考えられる．新語の追加は非常に時間がかかる為，本研究では “紙” の一項目のみの実験にとどめた．“紙” という語は EDR 電子化辞書に既に存在する．つまり，EDR 電子化辞書に存在する “紙” と新語の “紙” のシソーラス上の位置を比較する事を評価基準とする．

google で “紙” を検索し，結果の上位 10 サイトを基サイトとして得た．この 10 サイトのアンカーを参照して “紙” という語が存在する文を収集した．また “紙” という語がページ内に存在する場合アンカーをたどり，再帰的にページを取得した．ここでは最大 5 ホップ先の URL まで “紙” が存在するページを探し，入力文を回収した．結果としておよそ 600 の入力文を得た．

これにより得た文を JUMAN で形態素解析を行ない，KNP で構文解析を行なった．格構造選択及び概念選択はスクリプトで半自動的に行なった．入力文は多ければ多いほど良いと予測し，結果として合計 512 の関係を入力した．入力例は表 2 の通りである．

8. 結果

結果は表 3 の通りである
“紙” の上位概念である，具体物，情報媒体などはリンクが強く，宗教，野菜，医療器具など全く関係ないものはリンクが弱くなっている．大まかには，“紙” の位置付けが出来ていると言える．しかしながら，“紙” と全く関係のない生命体のリンクがある．先出の例を見ると “紙を食べる” というタプルが存在する．これは，“山羊が紙を食べる” の様な例文が入力された為である．入力文の不足，片寄りにより間違った概念とのリンクが強くなる可能性がある事を示唆している．より多くより正しい “紙” に関する入力を Web から行なえば解消されると考えられる．

表 3: 結果

「紙」の上位概念	重み
具体物	0.99
機能で捉えた具体物	0.93
情報媒体	0.85
生命体	0.76
書いたもの	0.76
紙	0.69
⋮	⋮
宗教	0.01
野菜	0.01
医療器具	0.01

9. 考察

9.1. Web 資源の有効性

Web 資源は巨大でありその情報量はとてつもなく大きい．しかし，その情報の正確性は保証されていない．本研究では Web 資源から新語の自動学習を行なった．512 関係という入力は充分であるとは言えないが，本研究の実験結果を見るとある程度の有用性，信用性があると言える．1 つの新語について入力文の数がどの程度必要であるかを今後考察しなければならない．また，他の様々な語で実験する必要があると言える．今回新語とした “紙” は一般語であるが，専門語の新語である場合には違った結果が予想される．

9.2. 記憶容量，速度について

語やその意味は open-end であるため莫大な記憶容量が必要となる可能性がある．再構築は，入力 1

表 2: 入力例

検索文字	概念識別子	共起関係子	関係識別子	共起動詞	概念識別子
紙	3c1038	を	object	作る	0fe812
紙	3c1038	で	source	作る	0fe812
紙	3c1038	を	object	揃える	0fb435
紙	3c1038	が	object	汚れる	3d1aca
紙	3c1038	が	object	生む	0f5ecf
紙	3c1038	へ	object	デザインする	0ffa98
紙	3c1038	を	object	接着する	3cf0fd
紙	3c1038	と	goal	呼ぶ	061c7d
紙	3c1038	を	object	食べる	3bc6f0
紙	3c1038	と	goal	呼ぶ	061c7d
紙	3c1038	を	goal	糊づけする	10337a
紙	3c1038	を	object	組み合わせる	3cef7f

件に対して数分かかる．よって，今回のような数百件の入力ならば問題ないが，コーパスなどからの自動入力などは長時間になると考えられ高速化が課題となる．これらの問題にはリンクの重み付けの手法やリンクの消し方，不用な単語の削除のアルゴリズムの最適化が必要である．

10. おわりに

Web を入力として意味空間の再構築を行なったが，新語の位置がおおよそ推定できている．しかし，完成度はまだ低く改善の余地が多くある．Web の自動検索方法，リンクの強化の手法やリンクの消し方の改善や入力文からの意味解析の自動化などまだまだ課題は多い．入力を Web 全体から自動で高速に行なうことが出来れば広い知識をこのシステム上に乗せることができるのではないかと考えられる．

参考文献

[1] 電子化辞書仕様説明書 第 2 版： 日本電子化辞書研究所 (1996)

[2] S. Bechhofer, F. V. Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, L. A. Stein： OWL Web Ontology Language Reference： W3C Candidate Recommendation(2003)

[3] Knowledge, Information and Data Processing Group. : WordNet.OWL : http://taurus.unine.ch/GroupHome/knowler/wordnet.html

[4] 黒橋 禎夫, 河原 大輔： 日本語形態素解析システム JUMANversion4.0(2003)

[5] 黒橋禎夫： 日本語構文解析システム KNPversion2.0b6 使用説明書 (1998)

[6] C.J.Date： An Introduction to Database Systems： Addison-Wesley Pub(2003)

[7] 益岡隆志, 田窪行則： 基礎日本語文法ー改訂版ー： くろしお出版 (1992)

[8] 長尾真： 岩波講座ソフトウェア科学 14”知識と推論”： 岩波書店 (1988)

[9] 長尾真, 他： 岩波講座ソフトウェア科学 15”自然言語処理”： 岩波書店 (1996)