5P-1

web 辞書との比較による EDR 辞書の更新支援システム Dicorret

深田　隆恭[†]　　佐竹　洋樹[†]　　松村　冬子[‡]　　原田　実[‡]

青山学院大学理工学部情報テクノロジー学科[†‡]

1. はじめに

原田研究室は SAGE[1][2]と呼ばれる意味解析システムを開発し、質問応答や要約などに用いている。SAGE の語意解析には EDR 電子化辞書[3]を使用している。しかし、この EDR 電子化辞書は新聞データに基づいて複数人によって作成されたので、１つの単語における同じような語意の重複や、語意頻度が社会通念と異なっているなど、意味解析の結果を誤らせる原因を含んでいる。図１はその解析誤りの一例であり、「買収する」の語意説明が「土地を買い取る」と限定的なものになっており、文章の意味が通らなくなっている。本研究では、こういった問題を解消するために web 上の辞書の語意との比較に基づいた EDR 電子化辞書の半自動更新支援システム Dicorret を開発した。

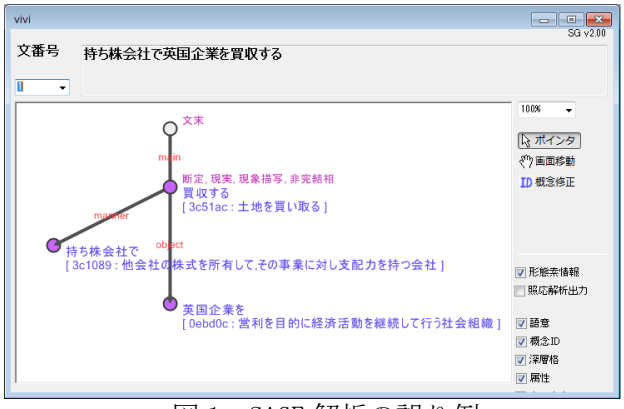

図１　SAGE 解析の誤り例

2. 提案システムの概要

Dicorret では図 2 に示す通り、まず修正すべき条件を持つ語を EDR 電子化辞書の中から検索し抽出する。そしてその語についての EDR 電子化辞書の語意と web 上の辞書に含まれる語意を取得し語毎に語意リストファイルにまとめる。

次にその取得したファイルに含まれる語の語意説明文を意味的類似度に基づいてクラスタリングを行う。しかし、機械によるクラスタリングは必ずしも正しいわけではないので、目で見て確認し間違っていれば人手で修正する。

クラスタリングが完了するとその結果に基づき EDR 電子化辞書を更新するための辞書更新レコードを生成する。

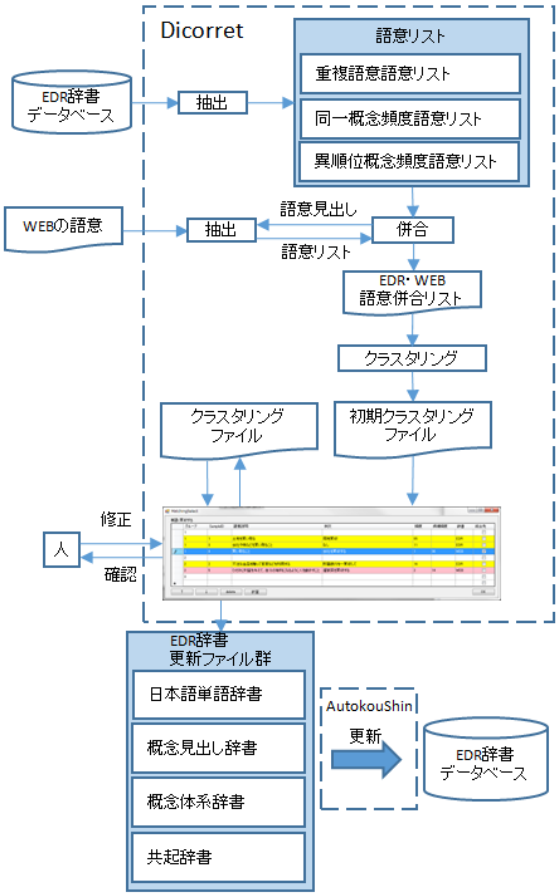

図２　Dicorret の概要

3. 提案システムの詳細

3.1 更新対象となる単語の抽出

EDR 電子化辞書から下記の三つの条件を持つ語を抽出する。また、リストに登録された語は web 辞書から語意説明文を取得する。

① 重複概念語意リスト

一語の中で語意説明文を対象に最小値型でリンク閾値を 0.3 としてクラスタリングを行う。一つのクラスタに EDR 辞書からの語意文が二つ以上存在するとき、その語を重複語意語意リストに登録する。

② 同一概念別頻度語意リスト

一語の中で概念別頻度が等しい語意説明をもつ語を同一概念別頻度語意リストに追加する。

③ 異順位概念別頻度語意リスト

Semi-Automatic EDR dictionary corrector Dicorret by comparing with web dictionary.

†Undergraduate school of Integrated Information Technology, Aoyama Gakuin University.

‡Faculty of Science and Engineering, Department of Integrated Information Technology, Aoyama Gakuin University

EDR 電子化辞書にあるすべての動詞や普通名詞について、語意が 5 つ以上ある語の語意説明文を対象にリンク閾値 0.2 でクラスタリングを行う。その結果 EDR 電子化辞書での語意頻度の順位と web の辞書で表記されている順位に違いがあればそれを異順位概念別頻度語意リストに登録する。

3.2 語意のクラスタリングによる重複語意の統合

3.1 節で取得した EDR 電子化辞書および web 辞書上の語意説明文を対象に語の階層型クラスタリングを行う。しかし、実際には web 辞書の語意説明と EDR 電子化辞書の語意説明は同じ内容を表記しているがその説明文が大きく異なる場合が多々あり、機械によるクラスタリングだけでは正しいクラスタリングができるとは限らない。そのため最終的にはクラスタリング結果を GUI に表示し、クラスタリング結果に問題がある場合は人手で修正する。例えば図 3 に示すように修正することで、web の語意「買い取ること」に対して EDR 辞書の「土地を買い取る」と「会社や株などを買い取ること」の二つが対応して、EDR の語意に重複があることが分かる。クラスタリングが完了したら重複語意の場合はどの語意を残すかを人間が判断して決定する。また、語意頻度がおかしい場合は WEB 辞書の語意説明の順番を参考に EDR 電子化辞書の概念別頻度の更新値を決定する。この結果に基づいて EDR 辞書の更新レコードを出力する。

3.3 辞書更新レコードの出力

重複語意は一つを残して残りは削除する。そのため、まず日本語単語辞書及び概念見出し辞書から該当の語意を削除する更新レコードを生成する。また、削除する語意の上位概念を残す語意の上位概念として追加登録するため概念体系辞書の更新レコードを生成する。さらに、削除する語意を含む共起事例を残す語意に置換する共起辞書更新レコードも生成する。表 1 はその一例であり日本語単語辞書から重複語意を削除するためのレコードである。語意頻度を修正する場合は日本語単語辞書だけを修正すればよいので日本語単語辞書更新レコードだけを生成する。最後に EDR 更新ツールを用いて更新レコード群に纏められた変更点を EDR 電子化辞書に反映する。

表 1　日本語単語辞書の更新レコードの出力例

更新区分	レコード番号	辞書見出し	読み	不変化部	品詞
削除	JWD0589152	買収する	バイシュスル	買収	JVE
	品詞日本語表記	概念ID	語意説明	概念別頻度	単語別頻度
	動詞	202678	会社や株などを買い取ること	96	96

4. 結論

図 4 は修正後の EDR 辞書を用いた意味解析の結果であり、図 1 に対して「買収する」の語意が適切な語意「買い取ること」になっているのが確認できる。これにより本システムで EDR 辞書を更新する有用性が確かめられた。今後は、実際に多くの語意を修正する作業を行うこと、また新聞データなどの解析によってこの修正による SAGE 解析精度を調査する。

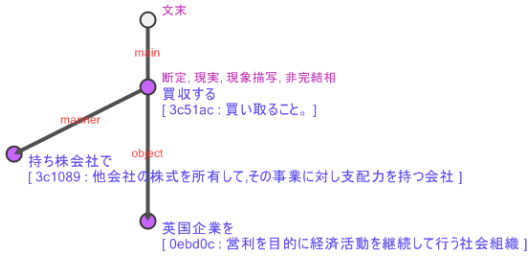

図 4　EDR 辞書修正後の SAGE 解析

5. 参考文献

[1]原田実, 水野高宏: EDR を用いた日本語意味解析システム SAGE , 人工知能学会論文誌, Vol.16, No.1, p.85-93 (2001).

[2]原田実, 田淵和幸, 大野博之, ″日本語意味解析システム SAGE の高速化・高精度化とコーパスによる精度評価″, 情報処理学会論文誌, Vol.43, No.9, pp.2894-2902 (2002).

[3](株) 日本語電子辞書研究所: EDR 電子化辞書仕様説明書 (第２版) ,(株) 日本語電子辞書研究所(1995).

[4]橋本広美,木下嵩基,原田実: ″フィルタリングのための隠語の有害語意検出機能の意味解析システム SAGE への組み込み″,情報処理学会研究報告, Vol.2010-NL-196, No.14, pp.1-6 (2010).

[5]安田 智成: 未知語に対する語意説明のインターネットからの獲得と電子辞書への自動登録, 修士論文, 青山学院大学(2005).

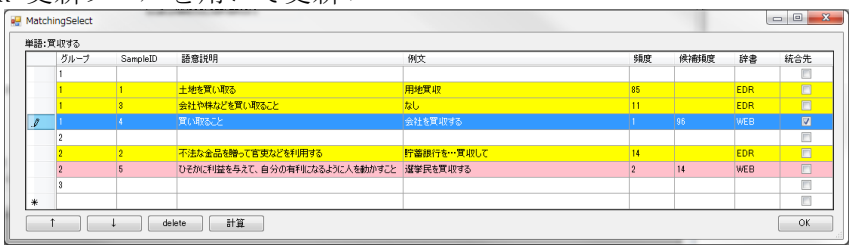

図 3　手修正後のクラスタリング結果