5Q-03

英文意味解析システム ESage

小林　豊†　　谷津元樹‡　　原田実‡

青山学院大学大学院理工学研究科†　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

意味解析とは文中の語意や語間の役割関係を解析することであり，質問応答やテキストマイニングや要約などの基礎技術である．原田らはこれまでに日本文の意味解析を行う日本語意味解析システム Sage [1]を開発してきた．本研究では Sage の技術を応用し，英文の意味解析を行う英文意味解析システム ESage を開発する．Sage と共通の係り受け構造・EDR 電子化辞書に基づく語意を利用して開発したので，日本文・英文間の概念類似度計算などへの応用が期待できる．

2. ESage の概要

ESage は The Stanford Parser [2]による係り受け解析の結果と EDR 電子化辞書 [3]の情報を基に意味解析を行っている．The Stanford Parser からは、係り受け情報以外に単語の品詞，原型，固有名詞かどうかの情報も利用している．ESage における処理の概要を図 1 に示す．

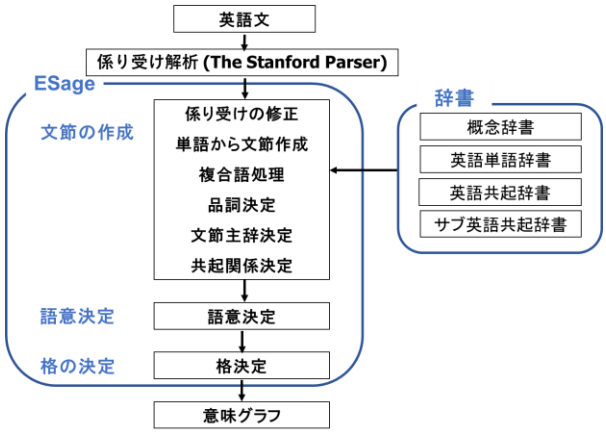


図 1. ESage のシステム構成

3. 辞書の構築

EDR 電子化辞書から必要な情報を抽出し新たにサブ英語共起辞書を構築した．この辞書は英語共起辞書に含まれる単語・共起関係子・品詞・概念情報を受け側，係り側に限定して集約し相対頻度(=該当レコードにおける共起項目頻度の和/該当レコード数)を追加した辞書である．

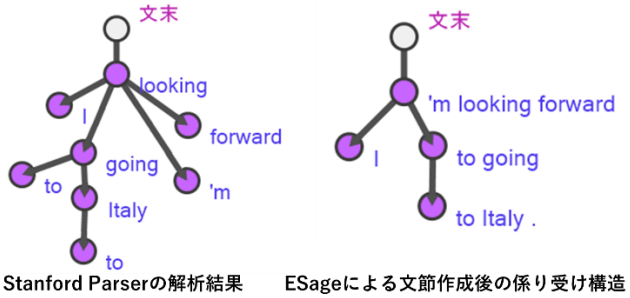
English semantic analysis system *ESage*
Kobayashi Yutaka† Yatsu Motoki‡ Harada Minoru‡
†Graduate School of Science and Engineering, Aoyama Gakuin University.
‡Faculty of Science and Engineering, Department of Integrated Information Technology, Aoyama Gakuin University.

4. ESage における意味解析

入力文の係り受け解析結果から，意味グラフとして意味解析結果を出力するまでの流れを以下に説明する．

4.1 文節の作成

Stanford Parser の解析結果は単語間の係り受けとなっている．係り受け情報および品詞情報を用いて単語を日本語と同様に文節としてまとめた．文“I'm looking forward to going to Italy.”の係り受け解析結果と文節作成後の係り受け構造の例を図 2 に示す．



Stanford Parserの解析結果　　ESageによる文節作成後の係り受け構造

図 2. 文節作成の前と後における係り受け構造

4.2 係り受けの修正

Sage による日本語の係り受け構造を参考に，係り受けの修正を行った．この修正により入力文によっては語順が変更される場合がある．

4.3 品詞・主辞・共起関係子の決定

単語，文節の品詞・主辞・共起関係子の決定について説明する．単語の品詞は Stanford Parser の品詞と EDR 電子化辞書の品詞の対応付を行い決定した．対応付けできないものはルールを作成し品詞の決定を行った．

主辞とは文節中で中心的な意味を示す単語である．共起関係子は，語間の役割関係を示す前置詞，接続詞もしくは記号である．どちらも 2 文節間の深層格を決定する際に辞書引きで用いる．主辞は優先順位を決め文節ごとに決定した．共起関係子はルールを決め品詞・係り受け情報・語順を基に決定した．

4.4 語意と格の決定

語意と格の決定について説明する．語意は約40 万種類の概念から，格は 37 種類から選択する．共起辞書から下記の 5 種類の方法で語意と格の候補を検索し出現確率に基づく評価を行い決定した．また冠詞の格など一意に決定できるものはルールにより決定した．

1. 共起関係子と 2 単語をキーとして検索（例：result of decision）
2. 共起関係子と 1 単語をキーとして検索（例：result of, of decision）
3. 共起関係子なし 2 単語をキーとして検索（例：result, decision）
4. 上位概念検索 （単語の上位概念を用いて総合評価を行う）
5. 単語辞書検索 （単語辞書から見出しで検索し語意のみ決定する）

4.5 意味グラフの出力例

文“Bills on ports and immigration were submitted by Senator Brownback, Republican of Kansas.”の出力結果を図 3 に示す．単語”bill”には「紙のお金」や「勘定計算の明細をしるした紙」などの概念もあるが，「法律の原案」という文意に合った概念を選択できている．また”of Kansas”という文節は place 格ではなく，修飾関係を示す modifier 格を選択できている．

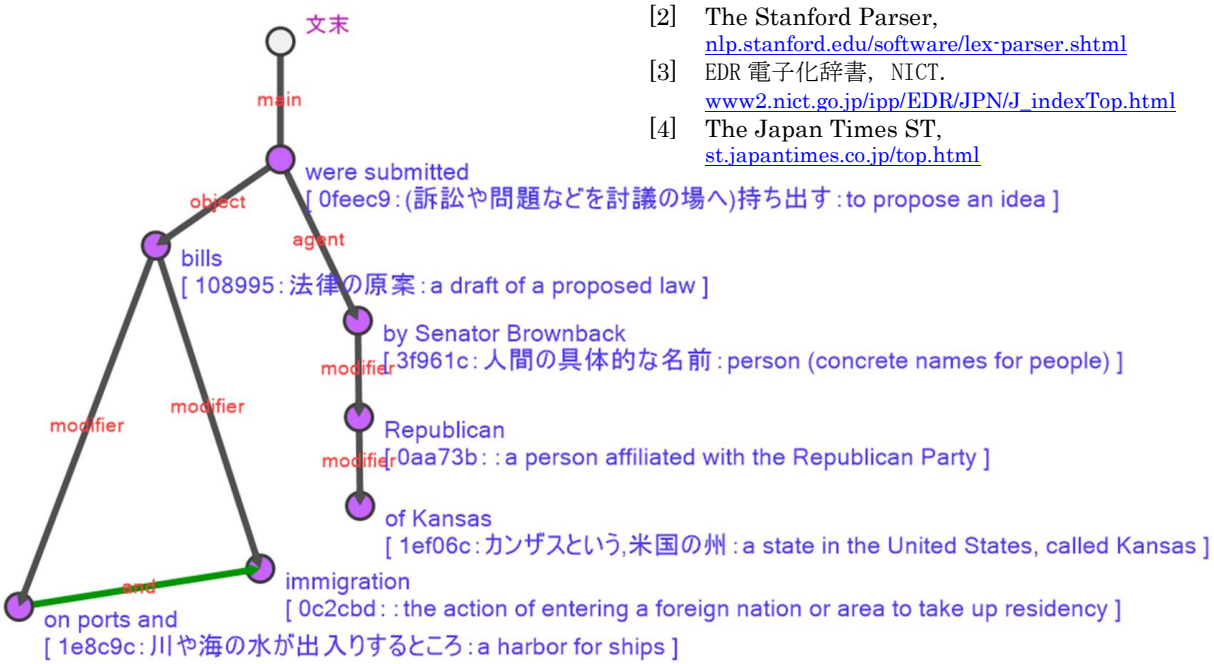


図 3. 意味グラフの出力例

5. 評価実験

The Japan Times ST [4]から 87 文を ESage で解析し，全単語と全係り受けの深層格における正解率(=正解数/全出力数)を人手で評価した．

結果を表 1 に示す．不正解となった語意の 48.0 %が固有名詞やハイフンで繋いだ複合語などの辞書に語意が登録されていないものであった．また，50 回以上出力された深層格で，正解率および F 値が最も高かったのは time 格（事象の起こる時間）であった．これは新聞記事に日にちなどを表す単語が多く，正確に語意を選択できているためである．

表 1. 評価実験結果

	語意の精度	格の精度
全出力数	1855	1095
正解率	86.8%	63.9%

6. まとめ

本研究によって 40 万を超える語意粒度において 86%以上の精度をもつ英文に関する意味解析システムを世界で初めて実現した．

今後の課題として深層格精度の低さが挙げられる．深層格は共起辞書や決定ルール，語意は辞書の充実により精度上昇が期待できる．

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, pp. 85-93 (2001. 1).
[2] The Stanford Parser, nlp.stanford.edu/software/lex-parser.shtml
[3] EDR 電子化辞書, NICT. www2.nict.go.jp/ipp/EDR/JPN/J_indexTop.html
[4] The Japan Times ST, st.japantimes.co.jp/top.html