Wikipedia の語彙ネットワークを用いたオープンドメイン Web 質問応答システム

065809　三枝　優一

（ 指導教員　速水　治夫　教授 ）

1.　はじめに

Web 上の文書情報を扱う既存のシステムとして**検索エンジン**がある．既存の検索エンジンは，本質的には利用者から入力されたキーワードに対し，そのキーワードを含む Web ページを利用者に提示するのみである．したがって，利用者は検索エンジンにより提示された文書の中から，求める情報を再度自らで探し出さなければならないという問題がある．

一方，Web 上には同種の情報が多様に表現されている．利用者は，どのようなキーワード入力が効果的に情報を絞り込めるのかキーワード選びに苦慮することが多く，多様な表現による複数回の検索を行なわなければならない問題がある．現状これは，利用者自身の知識や経験といった力量に依存している．

前者の解決策として，本研究では **Web 質問応答システム**を用いる．しかし，Web には自然言語処理に用いられる既存の言語資源には登録されていない多様な表現が数多く存在するため，解析ミスを引き起こしやすく，システム全体の精度低下を招く．そのため，獲得できる表現は情報の更新頻度が高い資源を利用して獲得することが必要である．

また，利用者の力量の問題に対する解決策として，1 回の検索で多様に表現される同種の情報を吸収し効率よく情報検索を行う手法を提案する．同種の情報を吸収するためには，同義語辞書や関連語辞書などの言語資源が必要である．しかし，Web 上には既存の言語資源には登録されていない多様な表現が数多く存在するため，獲得できる同義語・関連語表現は情報の更新頻度が高い資源を利用して獲得することが必要である．

そこで本研究では，人に優しい情報検索を実現することを目的とし，インタラクティブにキーワードを選定し，回答を直接提示する**インタラクティブ型オープンドメイン Web 質問応答システム**を提案する．これを実現するため，Web 上の多様な表現の獲得と，同義語，関連語関係を抽出する更新頻度の高い資源としてオンライン百科事典である Wikipedia[1]を利用し，Wikipedia シソーラスを構築した．

本稿では，Wikipedia シソーラスの構築手法，また Wikipedia シソーラスを用いた Web 質問応答システムへの拡張手法，ならびにそれぞれの有用性について評価した結果と考察について述べる．

2. Wikipedia シソーラスの構築

Wikipedia は，情報の網羅性・信頼性・更新頻度が高く一つのページが一つの用語を解説しているため，見出し語の解析ミスがないという特徴を持つ．また，リンクやリダイレクト，段落構造などの特有構造は新語，固有表現，同義語，関連語を抽出する手掛かりになる．Wikipedia シソーラスの構築関係を図 1 に示す．

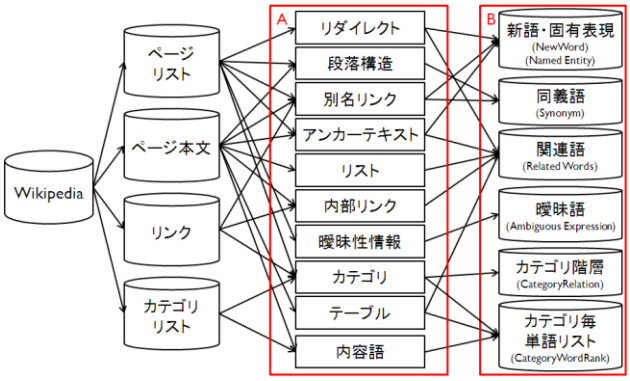

図 1 Wikipedia シソーラスの構築関係

初めに Wikipedia のデータベースから，”ページリスト”，”ページ本文”などの補助テーブルを生成した．生成した各補助テーブルから Wikipedia の特有構造を基に各処理(図中 A)を行い，”新語・固有表現”，”同義語”，”関連語”などのテーブルにそれぞれのデータを格納(図中 B)し Wikipedia シソーラスを構築した．

3. インタラクティブ型オープンドメイン Web 質問応答システム

一般的な質問応答システムは，”質問文解析部”,”情報検索部”,”回答抽出部”の 3 部からなる．これに対し本研究で提案するインタラクティブ型オープンドメイン Web 質問応答システムは Wikipedia シソーラスを適用し，質問文解析部と回答抽出部を拡張した．システム概要を図 2 に示す．

3.1 質問文解析部におけるキーワード拡張手法

Wikipedia シソーラスを適用した質問文解析部では，質問文中の内容語(名詞等)に関連語・曖昧語が含まれている場合，利用者は自らのイメージに沿う関連語やドメインをインタラクティブに選定し追加することで情報を絞り込むことができる手法を提案する．また，質問文中の内容語に同義語が存在する場合，内容語と併せて情報検索部への入力キ

ーワードとすることで情報の網羅性を向上させることができる手法を提案する.

3.2 回答抽出部における回答抽出手法

Wikipedia シソーラスを適用した回答抽出部では,頻度ベース手法としてナイーブ・ベイズ分類器を用いた機械学習によるカテゴリ推定手法を提案する.これは Wikipedia シソーラス構築時に学習したカテゴリ毎の単語リストの単語出現頻度から確率値を求め,ある文書がどのドメインに属しているのかを推定し,推定されたカテゴリ間の類似度を回答候補の重みに反映させ回答を抽出するものである.

また一方で,Wikipedia シソーラスに基づくベクトルベース手法として NP Vector を提案する.NP Vector とは,内容語(含 Wikipedia の新語・固有表現)と後続する格助詞とのペアを要素にしたベクトルである.この NP Vector を質問文と Snippet(検索エンジン返す縮約文書)において生成し,それらの類似度から回答を抽出する手法である.

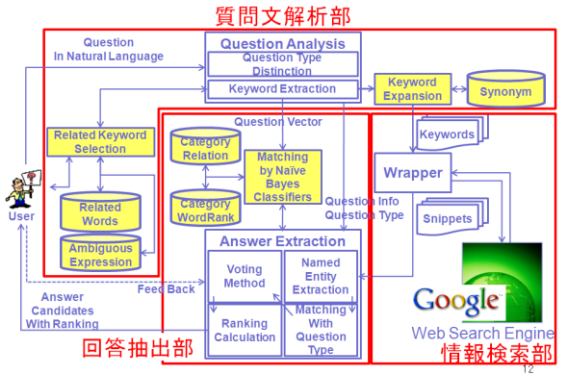

図 2 システム概略図

4. 評価実験

Wikipedia シソーラスとインタラクティブ型オープンドメイン Web 質問応答システムにおける提案手法の有用性について述べる.

4.1 Wikipedia シソーラスの評価

Wikipedia シソーラスで獲得した新語・固有表現数を図 3 に示す.我々の先行研究,既存の言語資源を代表する日本語語彙大系[2]より,Wikipedia シソーラスを用いた獲得語数が多いことがわかる.

先行研究	日本語語彙大系	Wikipediaシソーラス
24,802	約250,000	643,597

図 3 獲得新語・固有表現数

4.2 インタラクティブ型オープンドメイン Web 質問応答システムの評価

質問文のテストコレクションには NTCIR-4 QAC-2 をはじめとする 500 問を用い,キーワード拡張による同種の情報数(情報の網羅性)と回答抽出手法での正答数(回答精度)を集計した.

図 4 に,キーワード拡張の実験結果を,図 5 に回答精度の実験結果をそれぞれ示す.

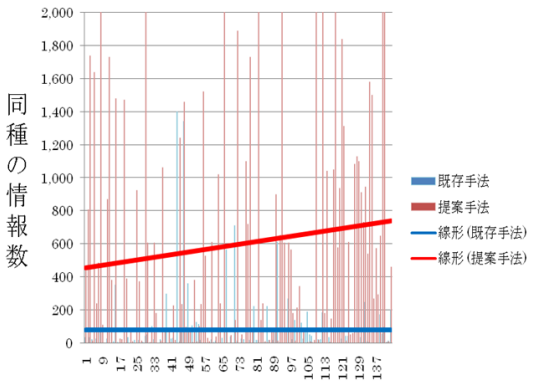

図 4 キーワード拡張実験結果

類似度比較 \ Wikipediaシソーラス		なし	あり
頻度ベース	PageRank なし	68 (従来手法)	81 (機械学習)
	PageRank あり	72	80 (機械学習)
ベクトルベース	PageRank なし	84 (Vector Space Model)	89 (NP Vector)
	PageRank あり	85 (Vector Space Model)	86 (NP Vector)

(単位:正答数/100問)

図 5 回答精度実験結果

図 4 より,本研究の提案手法は既存手法に比べ同種の情報数を約 6 倍多く獲得できることが確認できる.また,図 5 より本研究の提案手法である Wikipedia シソーラスを用いた手法は従来手法より回答精度が向上していることが確認できる.

5. おわりに

既存の検索エンジンにおける 2 つの問題点に対して本研究では「(利用者の力量を補うため)1 回の検索で様々に表現される同種の情報を吸収する」,「利用者が求める情報を直接提示する」という目的を設定し実験を行った.

「1 回の検索で様々に表現される同種の情報を吸収する」という目的に対して,Wikipedia シソーラスを用いることで同種の情報を約 6 倍多く獲得することができ本提案手法の有用性を確認した.

また,「利用者が求める情報を直接提示する」という目的に対して,Wikipedia シソーラスを適用したインタラクティブ型オープンドメイン Web 質問応答システムは従来手法より回答精度が向上したことを確認した.

参考文献

[1] Wikipedia : http://ja.wikipedia.org/
[2]日本語語彙大系 CD-ROM 版. 岩波書店(1999)