テキストコーパスと用語辞書に関する研究
～ コミュニティ辞書の構築と利用 ～

A Study on Text Corpora and Term Dictionaries

相澤 彰子（Akiko Aizawa）

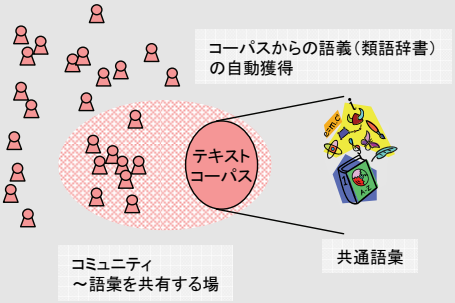
何がわかる？

特定の分野やコミュニティにおける語彙や用語遣いを反映する辞書を、コンピュータを使って簡単に作成できるようになります。これにより、分野に特化した情報の検索を容易にしたり、コミュニティ間での意味や用法の違いを比べたりすることが可能になります。

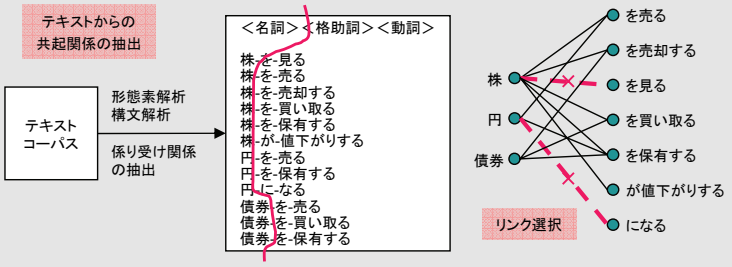
どんな研究？

Webや新聞記事など、手近に利用できる文書（「テキストコーパス」）を資源として活用して、コンピュータで「辞書」を自動的に生成する方法について研究しています。特に文書量の増加に伴う影響を調べ、類語や用例を網羅的に抽出する手法の実現を目指しています。

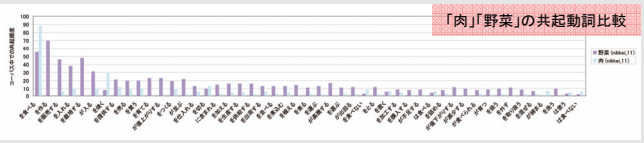
コーパス、コミュニティ、辞書



類語・用例辞書の抽出方式



語の頻度と類似度の影響



共起情報の抽出法:

（A）名詞・格付動詞
（B）名詞・動詞
（C）文内共起名詞
（D）名詞・格付動詞＋リンク選択

類似度の計算法:

(1) Jaccard尺度

$$sim_1(A, B) = \frac{V(A) \cap V(B)}{V(A) \cup V(B)}$$

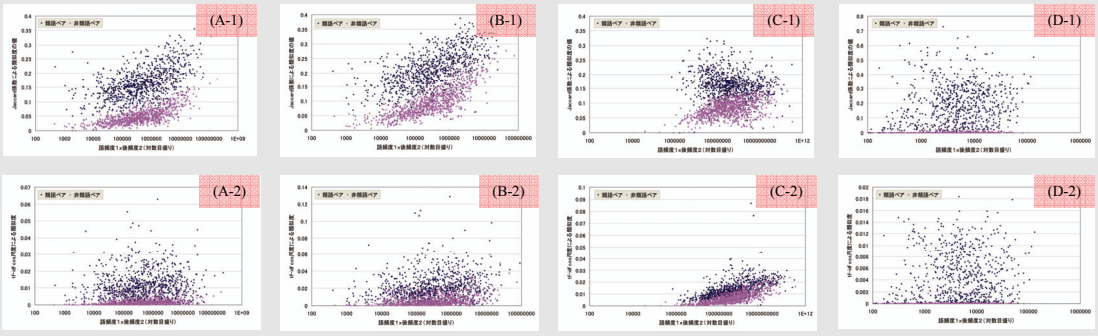
(2) tf-idf cos尺度

$$sim_1(A, B) = \frac{A \bullet B}{|A||B|}$$

評価方法:
対象コーパスから特定の表現パターンにより類語・非類語名詞ペアを抽出する

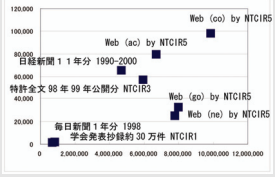
類語ペア
◇「A - や - B」という表記に注目して {A, B} を類語関係の候補とする
◇順序を逆にした {B, A} の出現頻度が極端に少ないペアを除外する
◇コーパスでの出現頻度がN＝5回以上、かつA, Bともに分類語彙表の見出し語で同じカテゴリに分類されているものを類語ペアとする

非類語ペア
◇分類語彙表の上でAと第二階層以下が異なり、かつコーパス中での出現頻度がもっともBに近い語C{A,C} を非類語ペアとする



自動抽出した辞書の出力例

適用例:コーパス規模（バイト数）と抽出した見出し語数



多義性の検出例

【タマ】 全コーパス

1.変化球, ゴロ, 速球, 初球, 直球, ストライク, フォークボール, フライ, スライダー, 魔球, 飛球, 真っすぐ
（～を打てる。～が投げられる。～を投げられる。～を投げ込む。～を打たせる。～を打たれる。）

2.石ころ, 鉄球, 包み, 珠, 銃弾, 弾丸, 手榴弾, サイコロ, オモリ, マシンガン, 積み木, 飛び道具, 電撃, フン, バレーボール, 鉄パイプ, 包帯, バズーカ, 魚雷, キューブ
（～を投げる。～を転がす。～が転がる。～を撃つ。～を弾く。～を発射する。～をはじく。～を握らせる。～が撃つ。～を喰らう。）

コーパスによる語義の違い

【株】 日本経済新聞

株式, 債券, 銘柄, 不動産, ドル, 国債, 銀行株, 金, ハイテク株, 同社株, 土地, 現物, 日本株, 円, 資産, 物件, 先物, 大型株, 原油, 通貨, 有価証券, マルク, 現物株, 投信, 自社株, 債権, 会員権, ガソリン, ファンド, 株券, 社債, ユーロ, 株式相場, 米国債, 手形, 米国株, 債券相場, …
（～を買い取る。～を売買する。～を手放す。～を買い戻す。～を物色する。）

【株】 特許（NTCIR3による）

菌株, 形質転換体, クローン, コロニー, 大腸菌, 形質転換株, プラスミド, 酵母, ハイブリドーマ, ベクター, 培養物, 変異株, 細菌, 菌, 宿主細胞, 細胞株, 微生物, 遺伝子, 細胞系, 菌糸, 組換え体, 線維芽細胞, 胞子, ファージ, 菌体, 該微生物, 陽性クローン, 培養細胞, 藻体, カルス, …
（～は寄託される。～を選抜する。～を培養する。～を選別する。～を接種する。）

コーパスによる語感の違い

【クーリングオフ】 Web全ドメイン

返品, 訂正, キャンセル, クレーム
（～を申し出る。～が効かない。～が利く。～を申し立てる。～に応じない。）

【クーリングオフ】 Coドメイン

強姦罪, 偽証罪, 傷害罪, 即時取得, 暴行罪, 詐欺罪, 名誉毀損罪, 監禁罪, 窃盗罪, 強盗罪
（～は適用される。～が成立する。～が適用される。～が適応される。～は成立しない。）

NII
連絡先： 相澤 彰子（Akiko AIZAWA）／ 国立情報学研究所 コンテンツ科学研究系 テキスト・言語メディアグループ
TEL : 03-4212-2524 FAX : 03-3556-1916 Email : aizawa@nii.ac.jp