普通名詞換言辞書の構築

電気系　山本研究室

09108085　山形　祐輝

換言処理

換言：ある表現を同義の別表現に変換する処理

「泥棒」→「空き巣」
「雷鳴」→「雷の音」

換言は言語処理の様々なタスクで用いられている

要約、質問応答、検索　etc.

背景・目的

換言を行うために換言知識が必要
換言の研究では以下の方法が一般的

- シソーラス、国語辞典の語釈文から収集
- コーパス、WEBの文中の関係から抽出

人による換言の知識が得られるような
汎用な換言辞書は作られていない

完全な人手により普通名詞換言辞書を構築

辞書の構築

換言対象は形態素解析器JUMANから抽出
→形態素辞書に含まれる普通名詞16,524語

- 対象語を作業者の考えだけで言い換える
- 換言は簡易な表現を意識する
- 内容語は2~3語程度に収める

例)「折り鶴」→「紙で折った鶴」

- 多義語でない限り、換言対は1つ

辞書の構築 ―作業基準―

- 換言は付与されているカテゴリに従う

例）「クラス」カテゴリ：組織・団体　→「集団」
　　抽象物　　→「階級」
「王冠」　カテゴリ：人工物−衣服　→「王がかぶる飾り」
　　×「瓶のふた」：カテゴリにそぐわない

- 以下の場合は無記入とする
  - 換言語を思いつかない
    「ストライク」「アルカリ」etc.
  - 元の語の意味が明確でない
    「村八分」「氏神」etc.

作業の効率を上げる
無理な換言を行わない

辞書の構築 ―作業結果―

換言対象：16,524語
　換言対　：16,153語
　無記入　：980語

- 換言を思いつかなかった語：310語
- 意味が分からない語：670語

クエリ拡張

クエリ拡張：与えられたクエリに関係するクエリの
候補を追加する処理
検索クエリ：検索のためのキーワード

例） 与えられたクエリ：「換言」
拡張クエリ　：「言い換え」　同義
「還元」　同音異字
「換言　意味」　共起頻度

換言とクエリ拡張

Ellen Mら[1]の研究

結果が一意に決まらない単語のクエリの拡張
→WordNetの同義語、上位語、下位語が有効

換言辞書≒同義表現を集めた語彙資源
→WordNetの同義語と同様、クエリ拡張に有効

構築した換言辞書の有用性をクエリ拡張にて示す

WordNet
プリンストン大学の認知科学研究所が開発、運営を行っている
一般に公開されている英語のシソーラス

[1]Ellen M、Voorhees. Query Expansion using Lexical-SemanticRelations.In 17th International Conference on Research and development in Information Retrieval (SIGIR'94). p61-69, Springer London, 1994.1.

辞書の評価

元クエリで獲得した文と拡張して獲得した文の内容語で類似度計算を行う
類似度計算はJaccard係数とSimpson係数を用いる

$$Jacc = \frac{|X \cap Y|}{|X \cup Y|} \qquad Simp = \frac{|X \cap Y|}{\min(|X|, |Y|)}$$

X：元クエリで獲得した文の内容語の集合
Y：拡張して獲得した文の内容語の集合

各クエリに対して一文対一文の総当たりで計算
平均をスコアとする

辞書の評価 ー使用データー

換言辞書：普通名詞換言辞書と用言等換言辞書

比較対象：日本語WordNet同義語データベースver.1.0

検索対象：毎日新聞2年分（1999、2000）

元クエリ：換言辞書とWordNetで見出し語となっている
普通名詞とサ変名詞の組み合わせ

日本語WordNet同義語データベース
独立行政法人情報通信研究機構（NICT）が開発、運営を行っている
一般に公開されている日本語WordNetの同義語の対を収録したもの

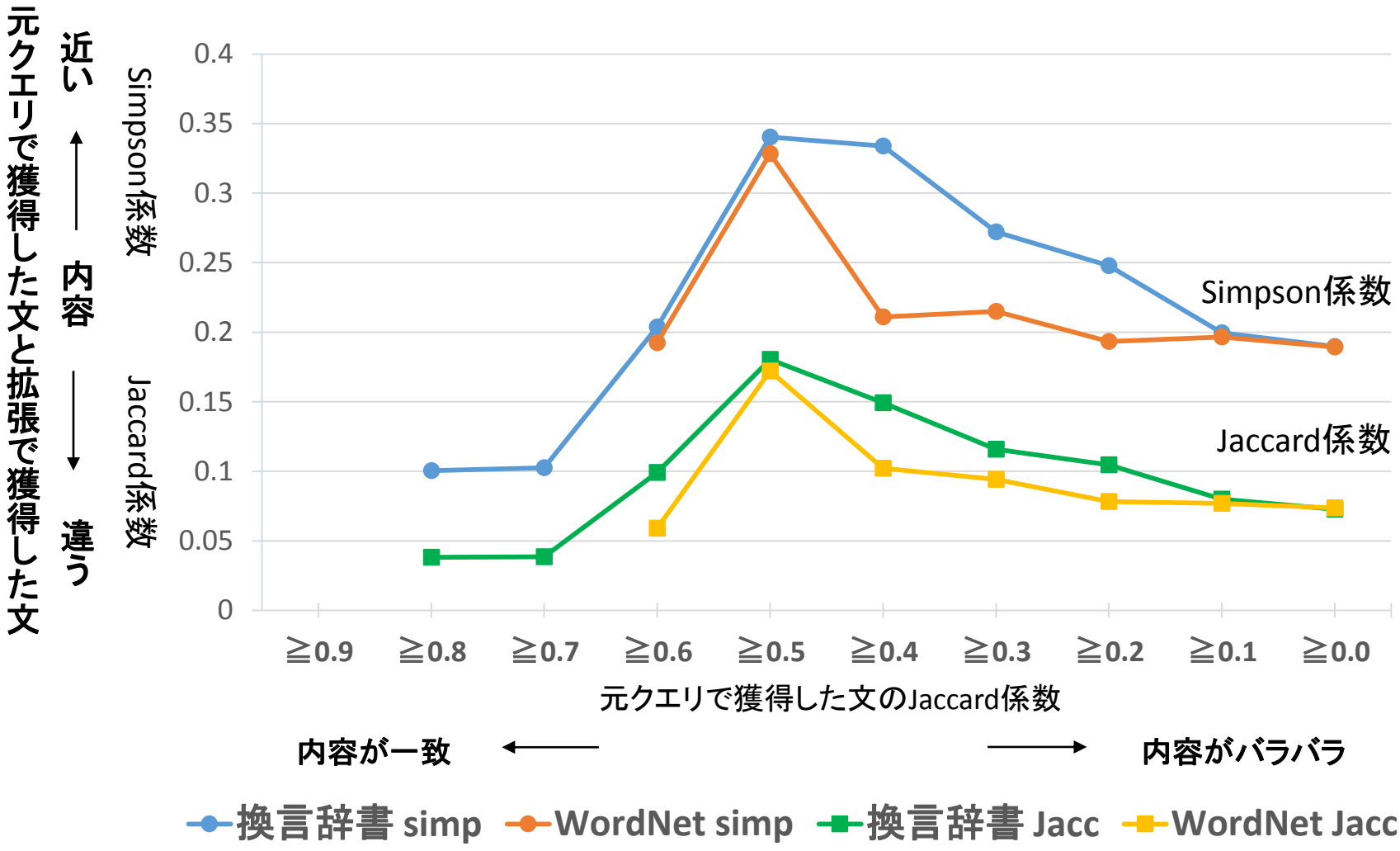
辞書の評価 ―結果―

	元クエリ	換言辞書	WordNet
クエリ数	24,510個	73,530個	1,074,212個
獲得文数	140,604文	＋110,237文	＋110,151文

換言辞書は文に出現しやすい語に拡張している

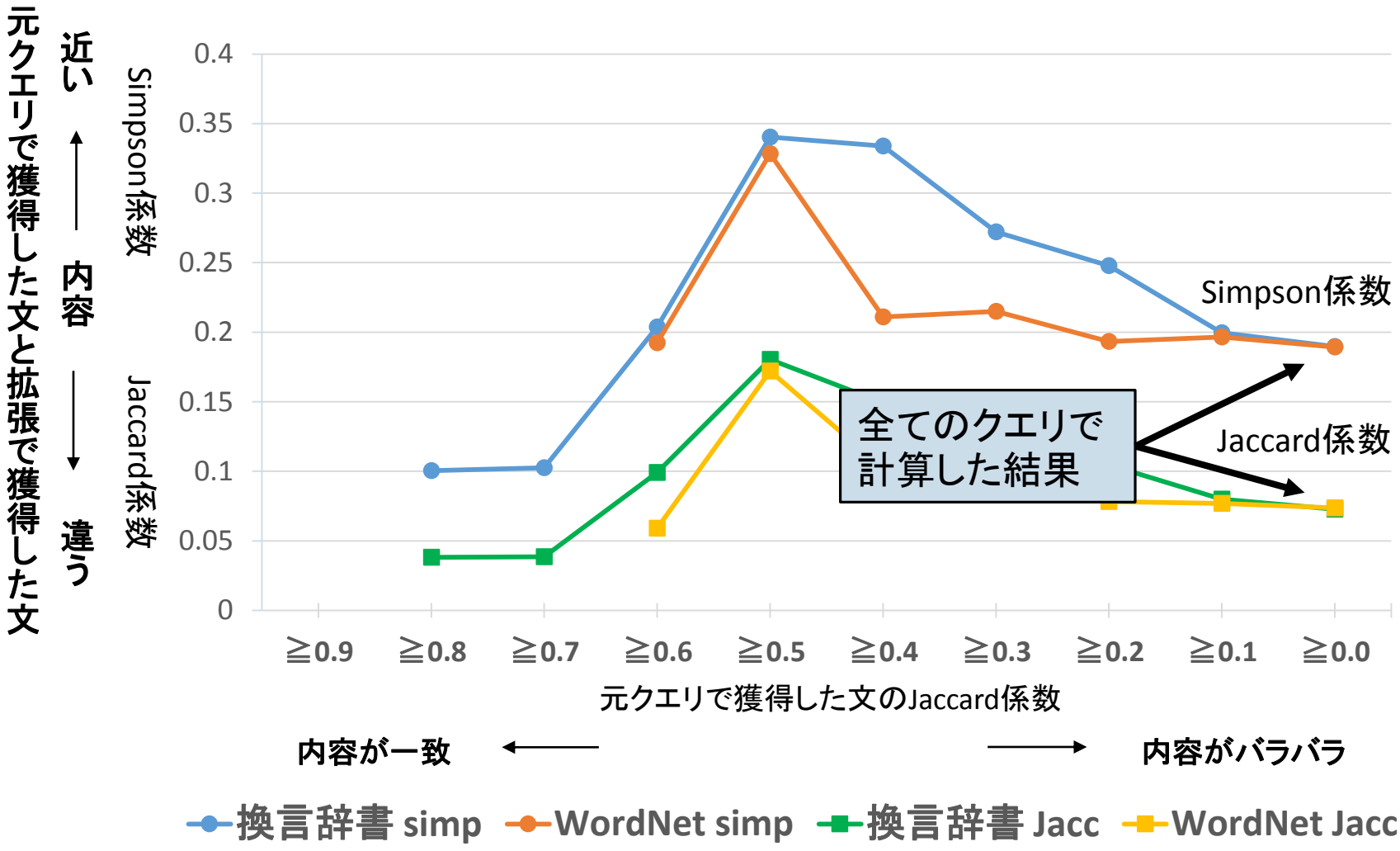
クエリ拡張の例） 元クエリ 「負債　削減」
換言辞書 「借金　削減」「負債　減らす」
「借金　減らす」
WordNet 「借入　削減」「負債　カット」
「負い目　カット」 etc.

辞書の評価 ―結果―



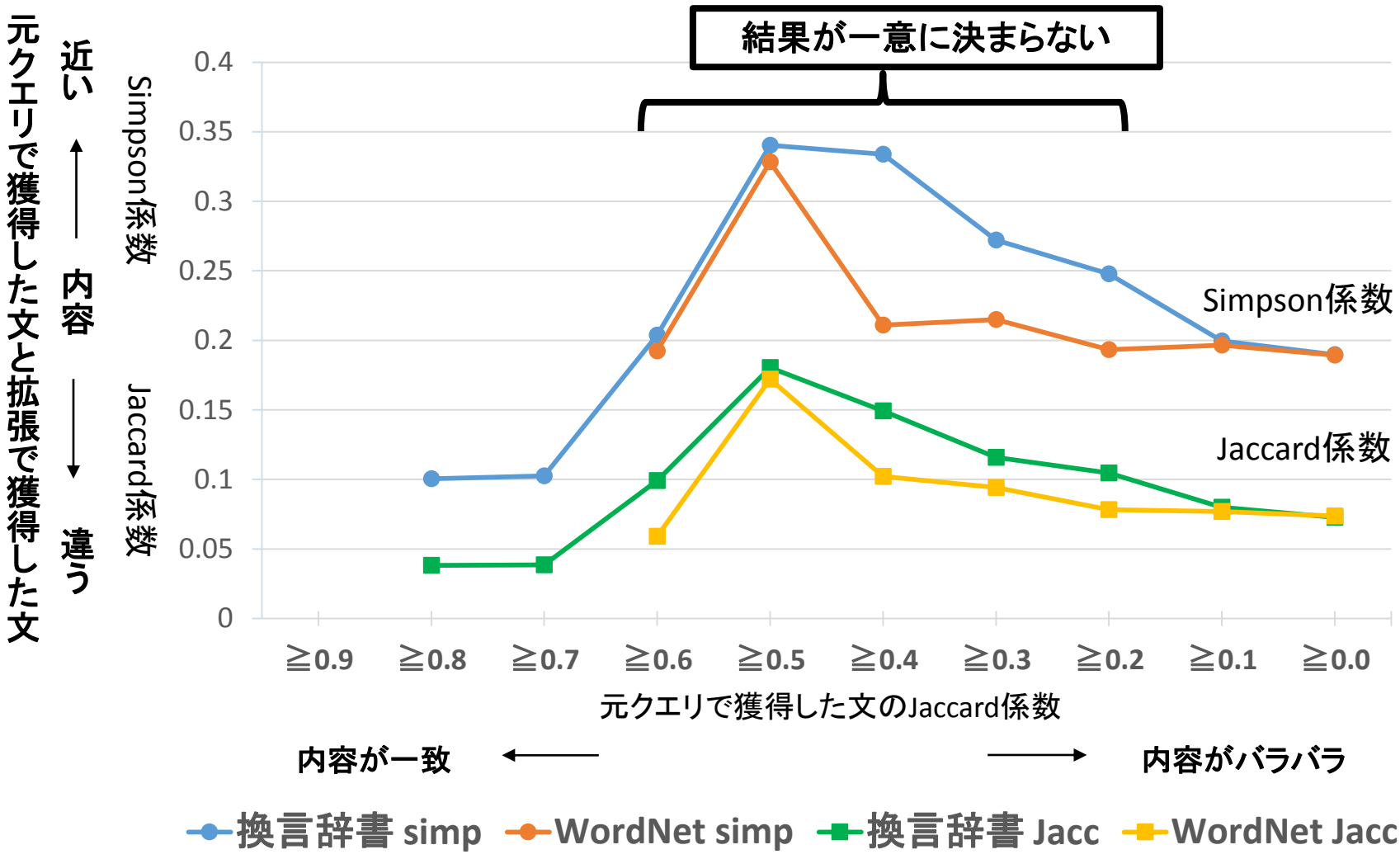
元クエリで獲得した文と各拡張で獲得した文の類似度計算結果

辞書の評価 ―結果―



元クエリで獲得した文と各拡張で獲得した文の類似度計算結果

辞書の評価 ―結果―



元クエリで獲得した文と各拡張で獲得した文の類似度計算結果

まとめ

完全に人手で普通名詞換言辞書を構築した

換言対象の約1万7千語に対し、約1万6千語の換言対を得た

構築した辞書の評価としてクエリ拡張を行った

普通名詞換言辞書と用言等換言辞書を合わせた換言辞書は
WordNetと同等以上の効果があることがわかった

換言辞書は公開する予定

ご清聴ありがとうございました

作業結果　詳細

カテゴリ	換言対象	換言作成	無記入
人工物	2,610語	2,557語	72語
自然物	453語	420語	33語
場所	1,795語	1,685語	111語
組織・団体	248語	228語	20語
人	1,479語	1,419語	66語
動物	771語	724語	47語
植物	339語	316語	23語
抽象物	6,912語	6,465語	435語
時間	259語	227語	33語
数量	353語	325語	29語
形・模様	135語	120語	15語
色	88語	84語	4語
複数	825語	1,583語	92語
合計	16,267語	16,153語	980語

換言辞書の拡張品詞別の結果

	普通名詞のみ	サ変名詞のみ	両方
Simpson係数	0.211	0.181	0.091

※サ変名詞の拡張
元クエリで獲得した文においてサ変名詞が用言として
使われている文とだけ計算

普通名詞のみ拡張した方がSimpson係数が高い