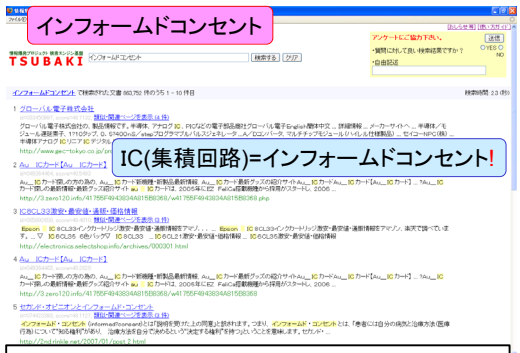
大規模コーパスに基づく同義語・多義語処理

京都大学 黒橋研究室　柴田知秀

◆ 自然言語処理において、同義語・多義語の扱いが常に問題となる
- 検索, 翻訳, 質問応答など

◆ 本発表の概要
- 国語辞典・Wikipedia・大規模コーパスから同義語・多義語を自動抽出
- 大規模コーパスを用いて、同義語のマージ、教師なし多義性解消

科研情報爆発で開発している検索エンジン基盤「TSUBAKI」において表現のずれを吸収



多義性解消を行なっていないため不適切なマッチングが生じてしまう

国語辞典・Wikipedia・大規模コーパスからの同義語・多義語・上位語の自動抽出

国語辞典からの同義語・上位語抽出

◆ 定義文のパターンで抽出
- 上位語
  -- 夕食: 夕方の食事
- 同義語
  -- アイス:「アイスクリーム」の略
  -- 購入: 買うこと （1文節）

◆高い網羅性で基本語彙の関係を抽出できるが、比喩的／例外的な関係も含む
- 犬:1/2 → 動物　0.353
- ~~犬:2/2 = スパイ 0.204~~
- 水道:1/2 = 上水道 0.362
- ~~水道:2/2 = 海峡 0.115~~

分布類似度の低いものを捨てる

大規模コーパスから分布類似度計算

◆「分布の類似した語は意味も類似している」[Firth 57]
◆ Web5億文から、係り受けに曖昧性のない用言・格要素を抽出
◆ 用言vと格cのペアを共起要素eと呼ぶ
- 例:「荷物を積む」→「積む:ヲ」が共起要素

◆ 名詞を共起要素のベクトルで表す
- 名詞と共起要素が相互情報量が正のものを利用

◆ 分布類似度: 共起要素の重複率

	荷揚げ:ヲ	搬入:ヲ	なる:ニ	届く:ガ	集散:ヲ		食べる:ヲ
荷物 = (	1,	1,	1,	1,	0,	…,	0)
物資 = (	1,	1,	0,	1,	1,	…,	0)

大規模コーパスから同義語抽出

◆ 括弧表現を利用
- ..A(B).., ..B(A).. → A=B

◆ 国語辞典からは抽出できない固有名詞・専門用語・新語の同義語を抽出できる
- 国際連合教育科学文化機関 = ユネスコ
- 大規模集積回路 = IC
- 大規模集積回路 = LSI
- 携帯電話 = ケータイ

分布類似度が高いものをマージ

Wikipediaからの多義語抽出

◆ 曖昧さ回避ページを利用

IC:1/7 集積回路
IC:2/7 インタークーラー
IC:3/7 インターチェンジ
IC:4/7 インフォームド・コンセント
…
IC:7/7リンパ球性脈絡髄膜炎

教師なし多義性解消

◆ 同義語（または上位語）をクエリとしてTSUBAKIから100件の文書を取得
◆ 同一文に出現する内容語を素性としてSVMモデルを学習

…半導体集積回路（LSI）や電子部品などの電子デバイスに関する…

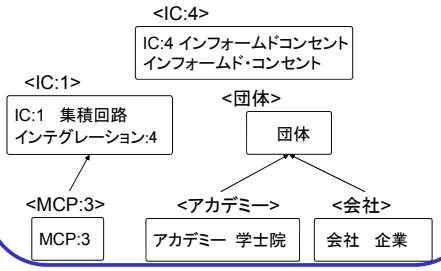
…春日インターチェンジは、京都府京都市西京区を通過する京都縦貫自動車道…

IC:1/7 集積回路 大規模集積回路 LSI
IC:2/7 インタークーラー
IC:3/7 インターチェンジ
IC:4/7 インフォームド・コンセント インフォームドコンセント
…
IC:7/7 リンパ球性脈絡髄膜炎

SYNGRAPHデータ構造　[Shibata et al. 08]

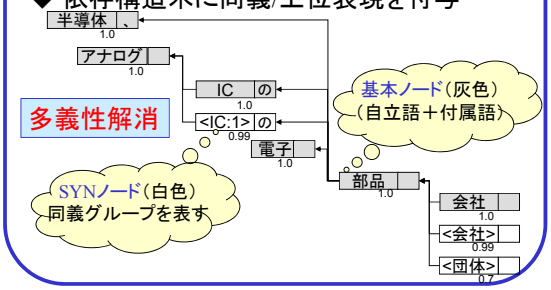
同義表現データベース

◆ 同義グループにSYNIDを付与



SYNGRAPHデータ構造

◆ 依存構造木に同義/上位表現を付与



まとめと今後の課題

◆国語辞典・Wikipedia・大規模コーパスから同義語・多義語を自動抽出
◆ 大規模コーパスを用いて、同義語のマージ、教師なし多義性解消
◆ 今後の課題
- 多義性解消を行なった同義語処理を検索エンジンTSUBAKIに実装
- NTCIRの検索コレクションで評価