日本語特許文書へのパーサー適応

橋本 力* 河原 大輔† 吉田 節行‡ 後藤 広樹§ 横山 晶一¶
*‡¶山形大学大学院理工学研究科 †情報通信研究機構 §山形大学工学部
{*ch@, §ecy72392@dipfr.dip., ¶yokoyama@}yz.yamagata-u.ac.jp
†dk@nict.go.jp ‡tma59172@st.yamagata-u.ac.jp

目 的 と 論 点

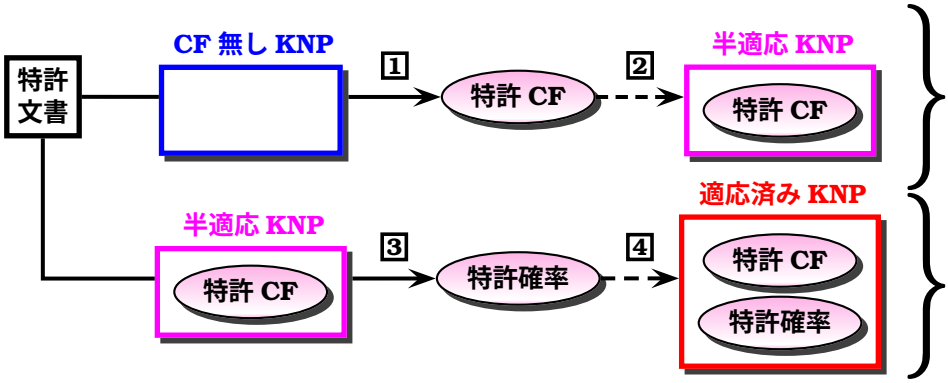
日本語特許文書へのパーサー適応

① どの適応手法がよいか？
- Ensemble 法
- Self-training 法
- 格フレーム法
- ···

② 異なる下位分野のデータは有効か？

例）生化学への適応に機械データは有効？
- 専門用語は違うが文体は同じ

格フレームによる適応法



1. 格フレーム（CF）無し KNP による特許文書からの特許格フレーム（特許 CF）の構築
2. KNP への特許 CF の組み込み
3. 半適応 KNP による特許文書からの特許格フレーム確率（特許確率）の計算
4. 半適応 KNP への特許確率の組み込み

格フレーム法の評価実験

実験条件

- ベースライン（BL）と適応済み KNP の精度を比較
  - BL： KNP ver 3.0, WebCF, Web 確率
- 対象ドメインとデータの規模（全て 2004 年のデータ）

ドメイン（IPC 分類）	CF 構築用	評価用
デジタルデータ処理（G06F）	3,500,000	1,000
微生物, 酵素等（C12N）	500,000	1,000
機械（F01）	250,000	1,000

- 文は 200 字以下のもの

実験結果

KNP	G06F 文書	C12N 文書	F01 文書
BL	0.893 ($\frac{10,644}{11,913}$)	0.893 ($\frac{9,340}{10,461}$)	0.902 ($\frac{11,890}{13,180}$)
G06F	0.896 ($\frac{10,677}{11,913}$)	0.893 ($\frac{9,345}{10,461}$)	**0.908 ($\frac{11,965}{13,180}$)**
C12N	**0.890 ($\frac{10,600}{11,913}$)**	**0.897 ($\frac{9,380}{10,461}$)**	0.903 ($\frac{11,897}{13,180}$)
F01	**0.890 ($\frac{10,608}{11,913}$)**	0.893 ($\frac{9,339}{10,461}$)	**0.909 ($\frac{11,979}{13,180}$)**

向上、有意　向上、非有意　低下、有意　低下、非有意

有意 ··· McNemar test, $p < 0.05$