D-72 統計調査の約600分類の符号付与システムについて

床 裕佳子, 下野 寿之, 和田 かず美, 坂下 佳一郎
（独立行政法人 統計センター）
2015/11/26 IBIS2015 つくば国際会議場

目的

文字列（家計簿記入内容）から

３桁の符号を高精度で正解を与える！

つまり

自然言語処理を活用して、

約５７０通りの分類を高精度かつ高速に行えるようなシステムを構築する。

データ

総務省統計局が実施する家計調査のデータ

データの件数

学習用：約１８７万件（約１９０ＭＢ）･･･１年分

符号付与用：約３２万件（約３０ＭＢ）･･･１ヶ月分

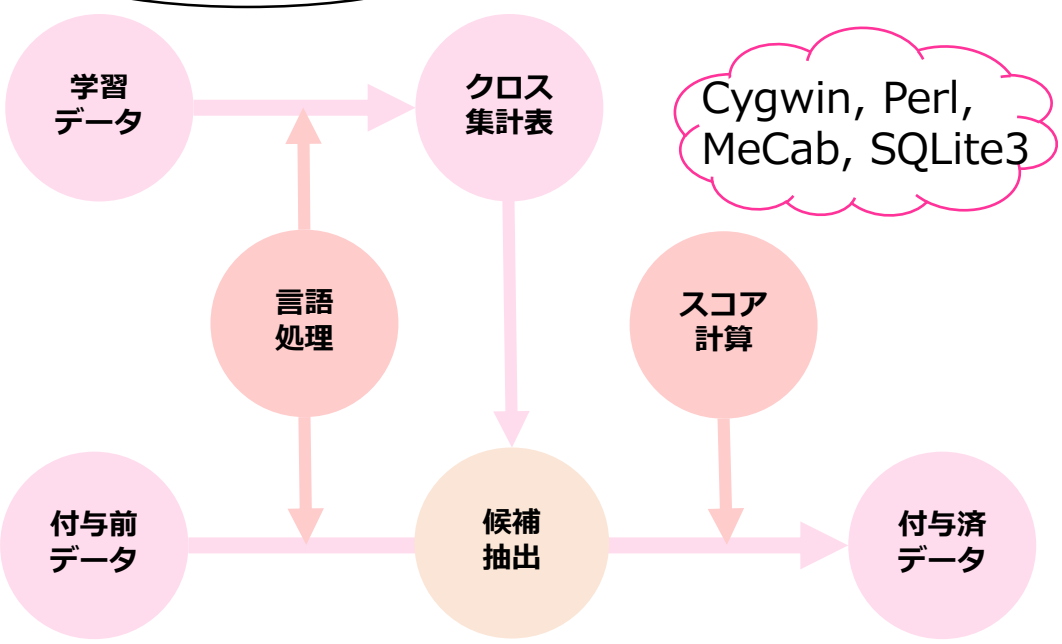
正解の分類符号

↓データイメージ

```
001 パン
548 カニカマ
313 給与（3月分）
281 にんじん2本
299 バナナ
411 ○○バーガーにて外食
002 チョコレートパン
328 チョコ
001 ロールぱん6ヶ入
088 キャベツ入りメンチカツ
315 所得税（6ヶ月分）
572 カニ
412 外食（うどん）
331 豆乳入りどら焼き
```

分類対象の文字列

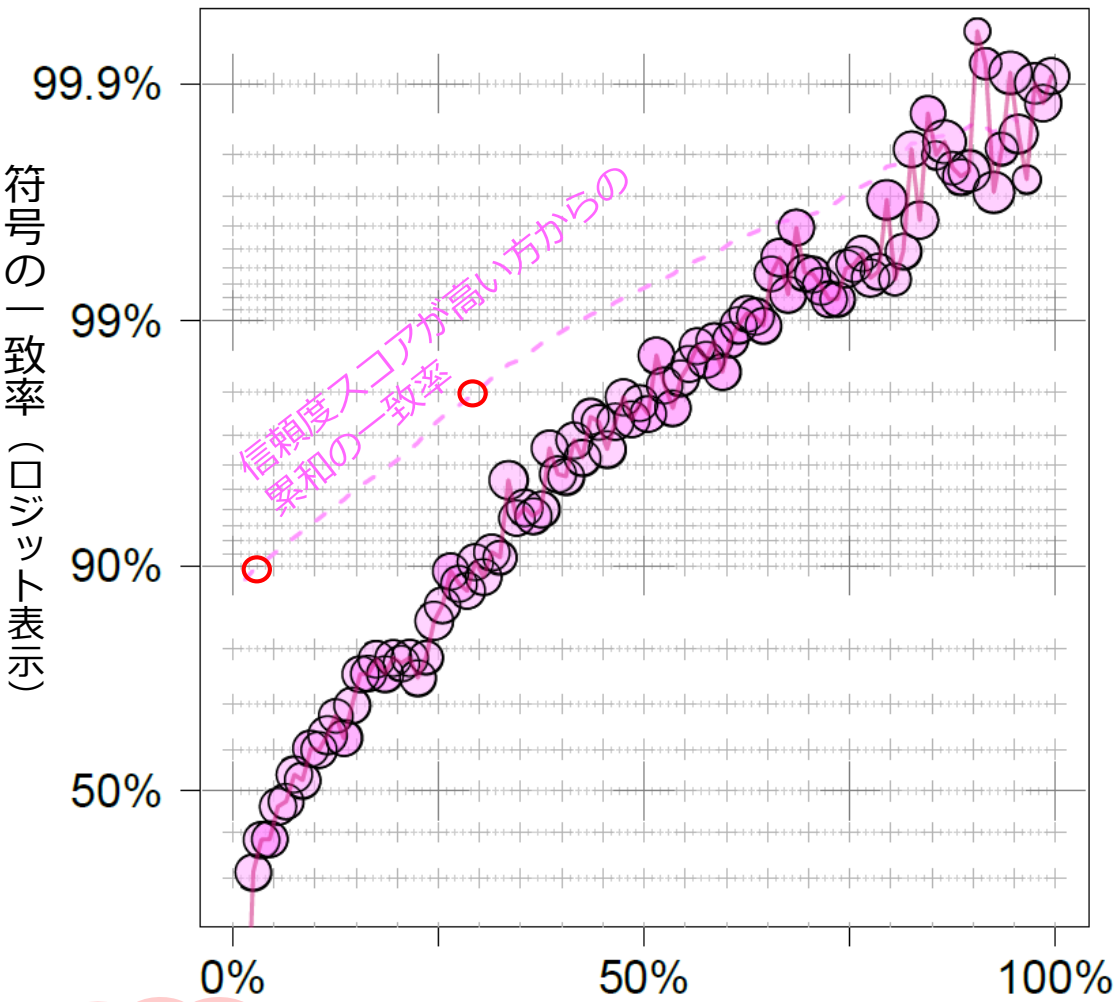
システム構成



Cygwin, Perl, MeCab, SQLite3

性能評価

信頼度スコアのパーセンタイルごとの一致率



全体の結果

一致率（一致件数 / 符号付与された件数） **８９．２%**

符号付与率（符号付与された件数 /全件数） **９８．５%**

さらに

一致率９０%保持ライン **９７．１%**

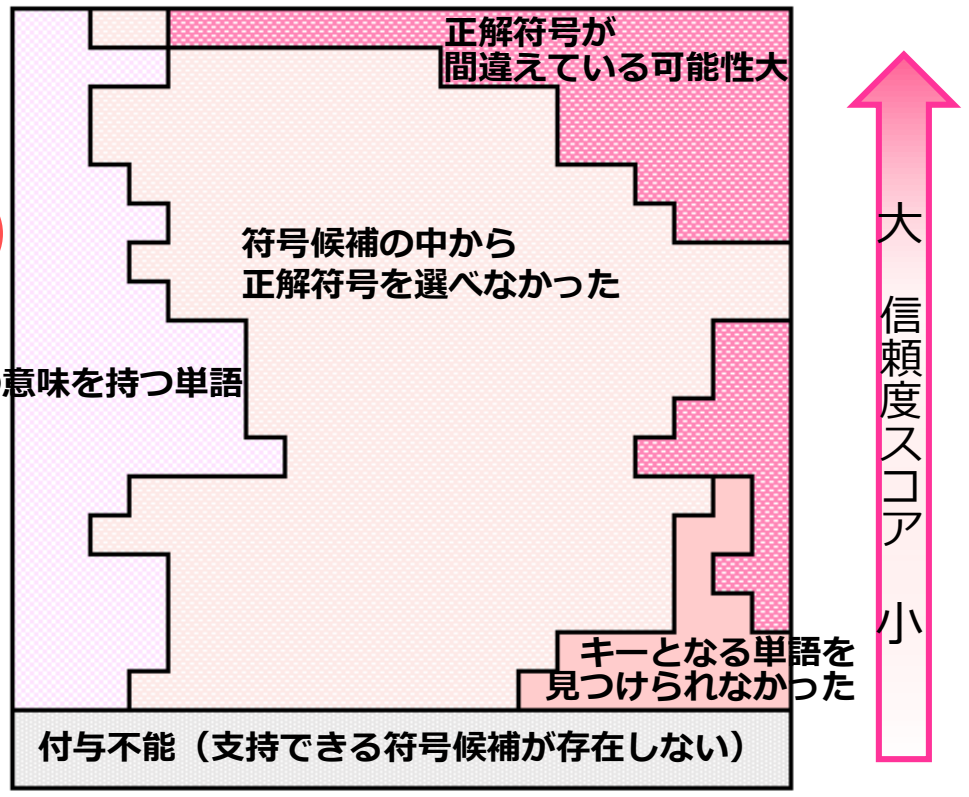
一致率９８%保持ライン **６９．６%**

処理時間（Xeon,3GHz）

学習フェーズ 約**２秒**/１万データ

符号付与フェーズ 約**５秒**/１万データ

不一致データの内訳（不一致データの件数で等積表示）



考察・課題

（１）**単純**･･･クロス集計表のみに依存
（２）**高精度**･･･専門的な知識を持った職員に近い精度
（３）**速い**･･･処理スピードが実用的

要改善ポイント

- 複数の意味をもつ単語
- ”強い”キーワードの副作用
- 表記ゆれ
- カタカナ連続記入 など

上記を踏まえ・・・**今後の課題**

- **他の属性**情報の活用⇒同音異義語/地域間表記ゆれ/収入か支出か
- **MeCabユーザ辞書**の拡充⇒カタカナ/表記ゆれ/記入ミスをカバー

注：本ポスターの内容は、報告者本人の見解であり、必ずしも所属組織を代表するものではありません。