テキストマイニングのための相互運用性と統合環境

誰でも使える言語処理を目指して

○狩野　芳伸
情報・システム研究機構　ライフサイエンス統合データベースセンター

入力テキストの選択

- BioNLP Shared Task 形式
- BIO/IOB 形式
- BioCreative 1a 形式
- 任意のテキストファイル
- Pubmed など公開文献

ユーザ側システムでの利用

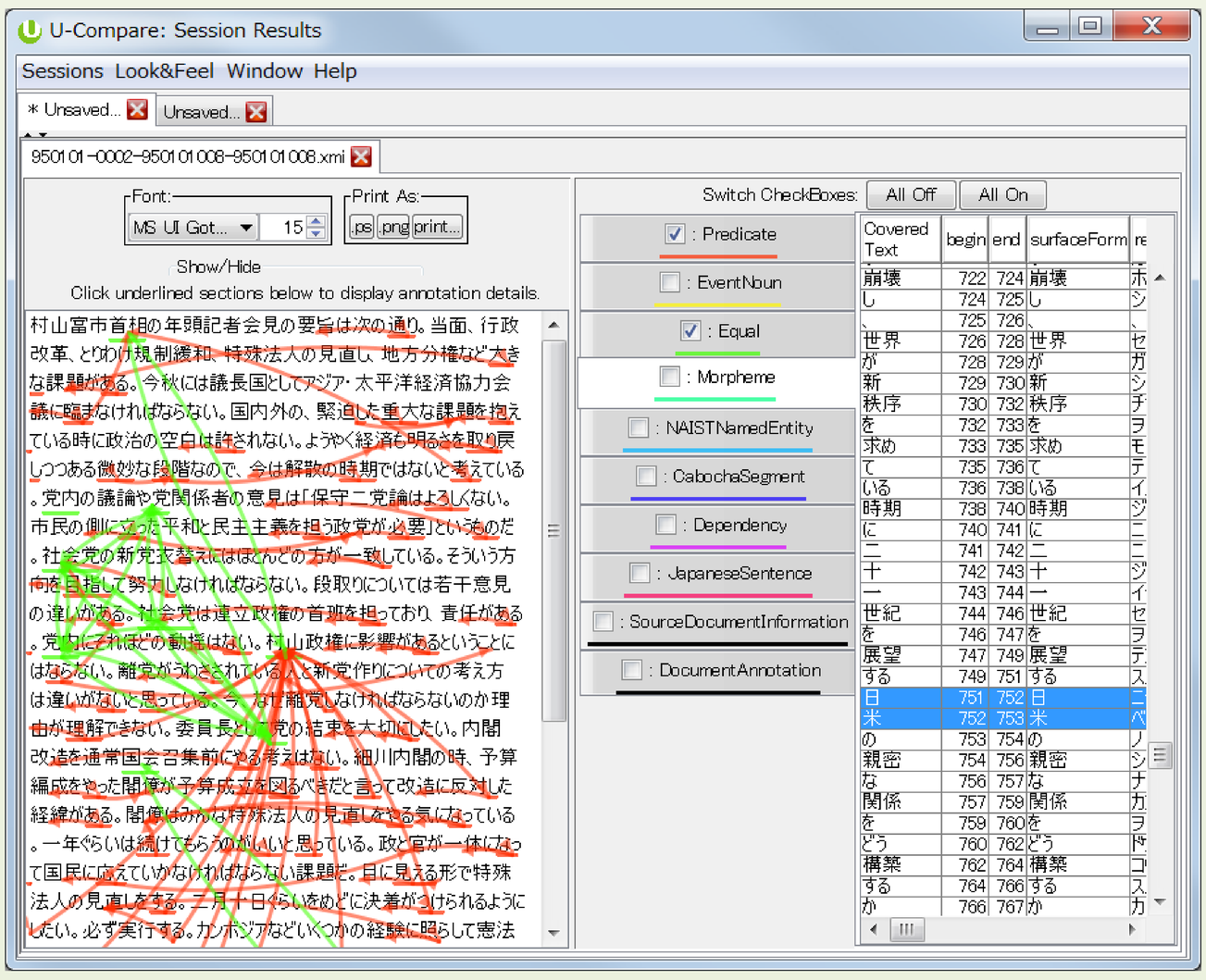
- コマンドラインからの呼び出し
- Taverna Plugin
- 指定フォーマットでの書き出し

テキストマイニング処理

U-Compare UIMAコンポーネント群:
即座に実行可能な世界最大の互換言語資源群

- コンポーネントの組合せ
- カスタマイズ結果の再利用
- ワークフロー作成
- 他ユーザとの共有
- 実行と結果解析で最適なフローを探索
- 他システムへの組み込み

視覚化・統計・自動組合せ比較・評価などの汎用機能



U-Compare
UIMA準拠 統合言語処理システム

本研究の概要

目的: 自動化と汎用機能
言語資源を最小の労力で利用可能に
汎用機能で典型的タスクをサポート

↓

必要なこと
言語資源の互換化・相互運用
インストール/組合せ/実行/解析 機能

↓

方法
国際標準準拠+共通データ型で統一
全自動の統合言語処理システム

↓

実装
生医学文献ツールのUIMA準拠・互換化
U-Compareシステムの視覚化機能
U-Compareへの言語資源統合

デモンストレーション
実行デモをお見せできますので
お問い合わせください

U-Compare とは

UIMA準拠の言語処理プラットフォームと互換言語資源群を統合したオープンソースのシステムです。即座に利用可能な言語資源群と、インストールから実行解析作業まで徹底した自動化により、言語処理タスクをプログラミング不要で完結させるのがねらいです。http://u-compare.org/ からいつでもお試しいただけます。

UIMA とは

UIMAとはテキストなど非構造化データのための相互運用性フレームワークです。実装はApache UIMAとして公開され、最近ではIBMのQAシステムWatsonで利用されるなど広く利用されています。

Database Center for Life Science
Licensed under a Creative Commons表示2.1日本 license
(c)2011 狩野 芳伸　科学技術振興機構
独立行政法人 科学技術振興機構 Japan Science and Technology Agency