Solrにおける日本語処理の現状

株式会社 ロンウイット

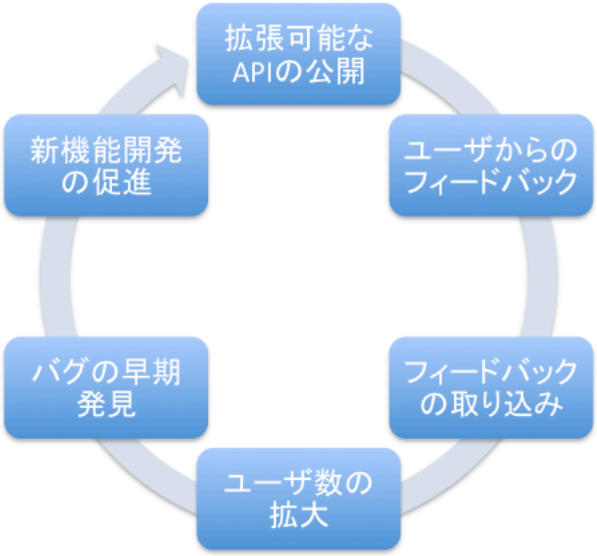
www.rondhuit.com

RONDHUIT
WWW.RONDHUIT.COM

Apache Lucene/Solr

- Apache Lucene
  - Javaで書かれたオープンソースの全文検索ライブラリ

- Apache Solr
  - Luceneをベースに開発されたオープンソースの全文検索サーバ



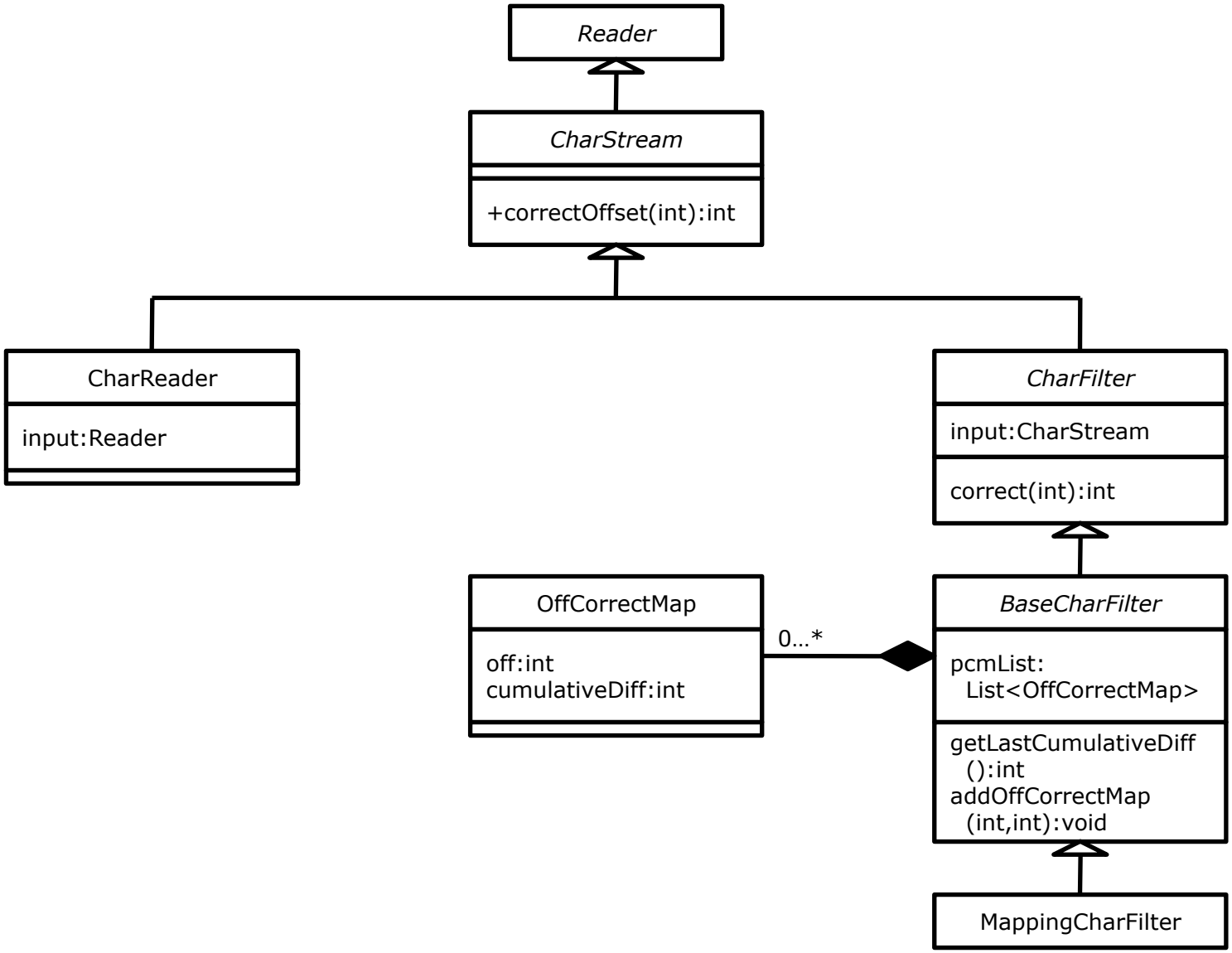
よくある質問／問題

- あいまい検索はできますか？
- 形態素解析とN-gramどちらを使えばいいですか？
- ハイライトが正しく動作しません。
- ハイライトがずれます。

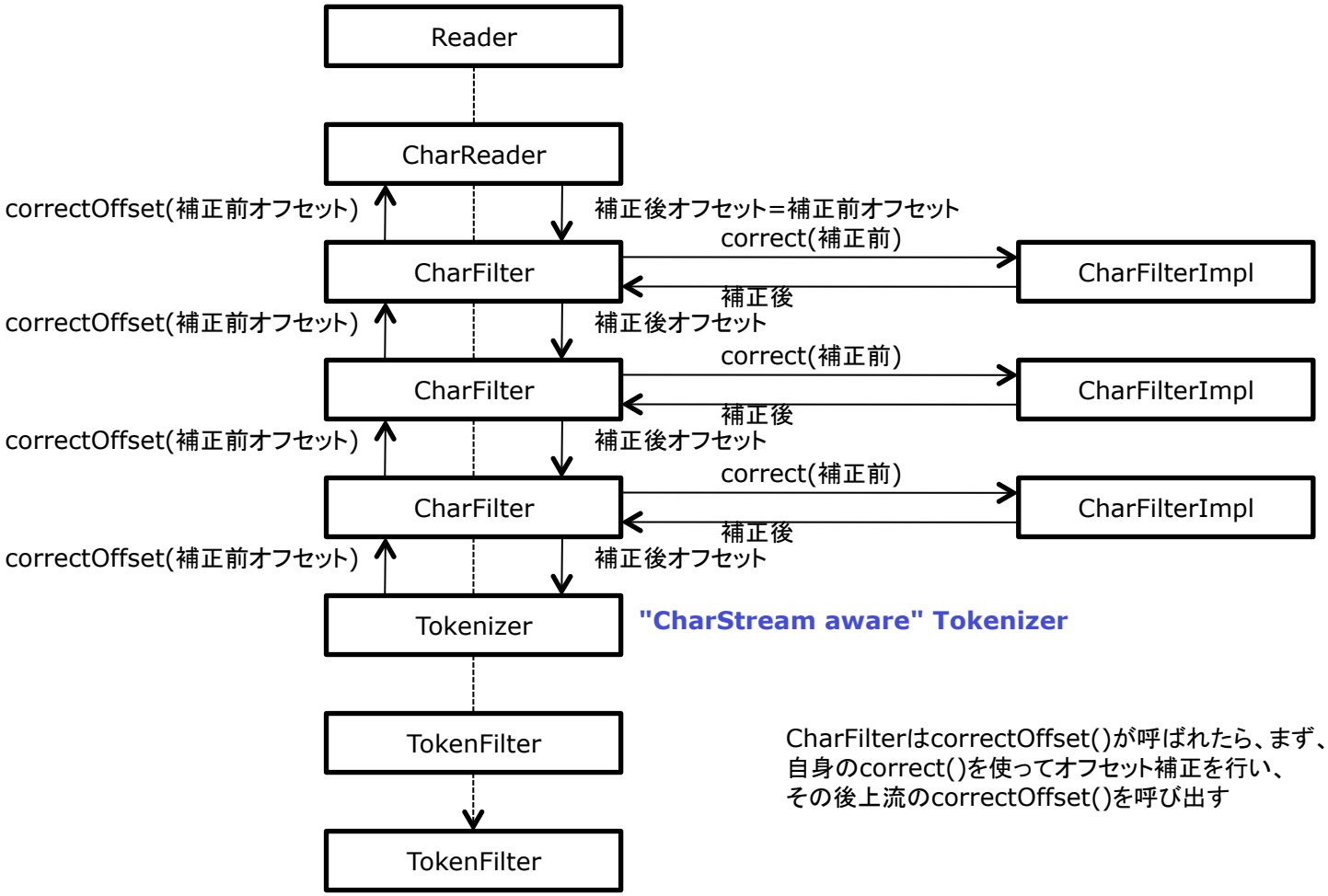
あいまい検索はできますか？

- あいまい検索とは？
  - ユーザによって「あいまい」の定義はさまざま
    - 全角・半角混じり文の検索
    - カタカナ語の表記揺れ
    - 正式名称⇔短縮名称検索
    - いすず⇔いすゞ
    - 結合文字検索
    - 西暦⇔和暦検索
    - ワイルドカード検索／正規表現検索／あいまい（Fuzzy）検索
    - Google「もしかして・・・」
    - Googleサジェスト

CharFilter – クラス図



文字オフセット補正の動き



CharFilterはcorrectOffset()が呼ばれたら、まず、
自身のcorrect()を使ってオフセット補正を行い、
その後上流のcorrectOffset()を呼び出す

MappingCharFilter

- <charFilter/>のmapping属性で指定されたファイルに従った「文字マッピング」を実行する。

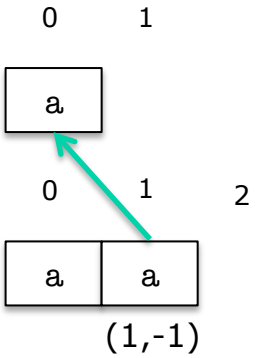
- mapping.txtのサンプル

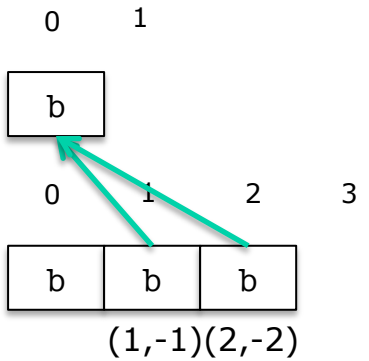
```
# Syntax:
#   "source" => "target"
#     "source".length() > 0 (source cannot be empty.)
#     "target".length() >= 0 (target can be empty.)

# example:
#   "À" => "A"
#   "\u00C0" => "A"
#   "\u00C0" => "\u0041"
#   "ß" => "ss"
#   "\t" => " "
#   "\n" => ""

# À => A
"\u00C0" => "A"

# Æ => AE
"\u00C6" => "AE"

# Ç => C
"\u00C7" => "C"
```

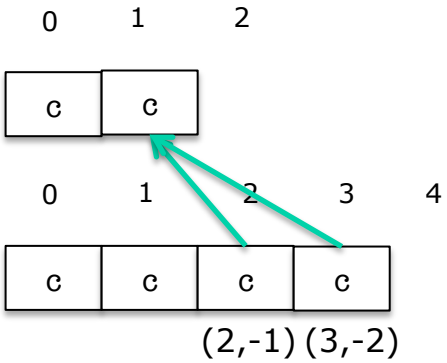
MappingCharFilter(オフセット補正)

a => aa
b => bbb
cc => cccc

文字数が増える場合



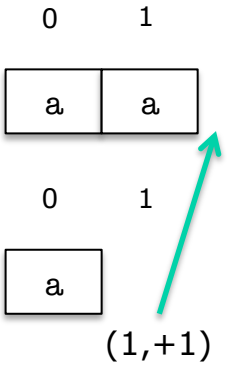


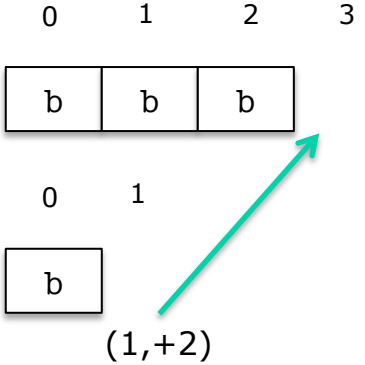


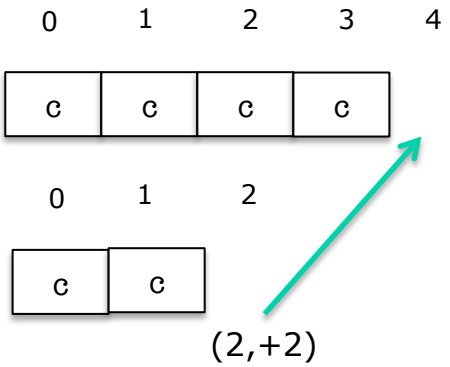
MappingCharFilter(オフセット補正)

aa => a
bbb => b
cccc => cc

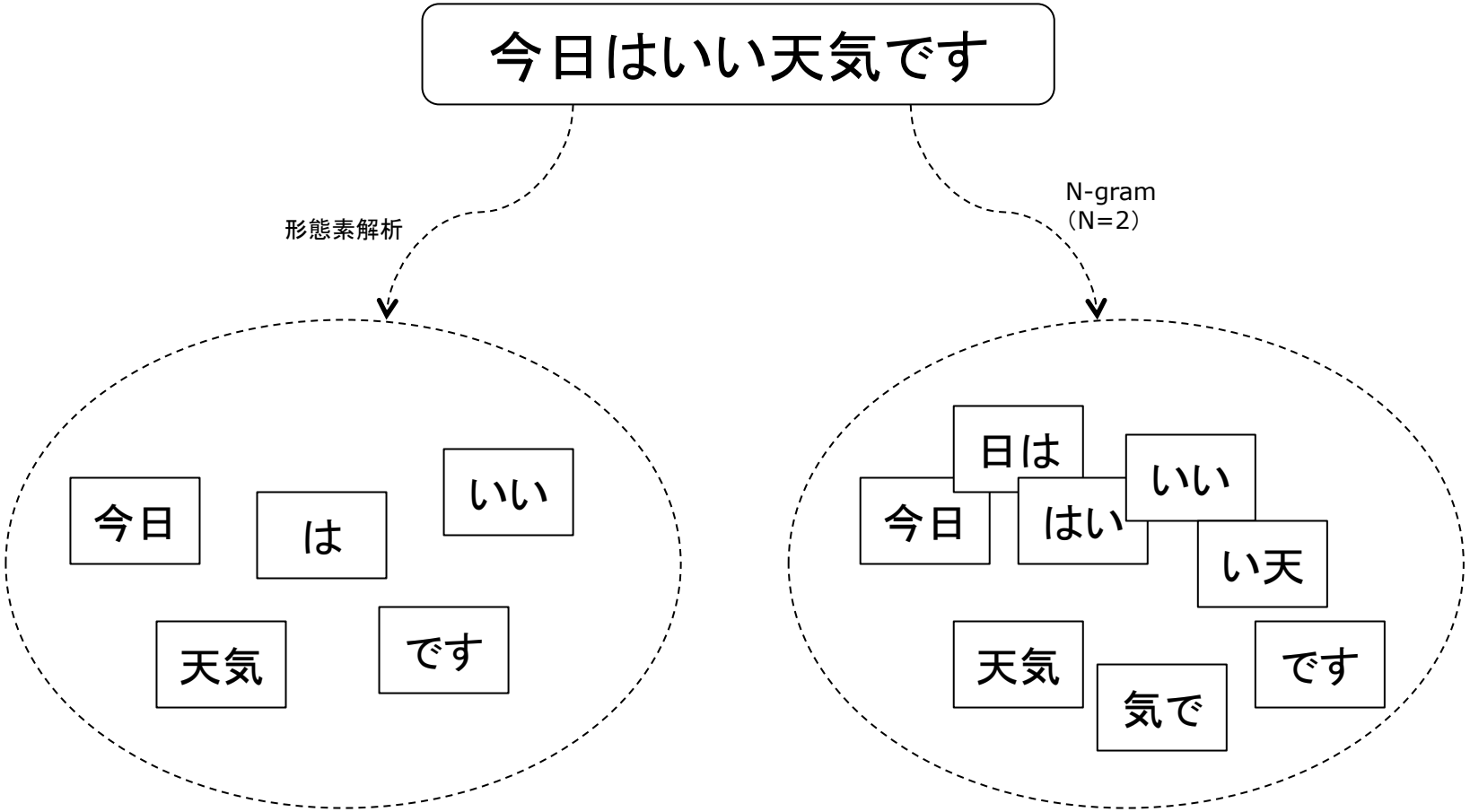
文字数が減る場合







形態素解析とN-gram



MeCab
Sen
Rosette言語処理プラットフォーム（RLP）

CJKTokenizer
NGramTokenizer

形態素解析とN-gram

	形態素解析	N-gram
長所	•自然言語の単語に近い •N-gramよりもインデックスサイズを小さくできる •単語の品詞が抽出できる	•実装が容易 •辞書不要 •新語への対応が容易
短所	•新語への対応が困難 •辞書のメンテナンスが必要 •N-gramよりも検索漏れが生じやすい	•形態素解析よりもインデックスサイズが大きくなりがち •Nより短い文字が検索漏れとなる
用途	•NLP •固有表現抽出	•特許検索 •ブログ検索

単語の境界がはっきりしない例

製造部門長谷川

ハイライトのトラブル

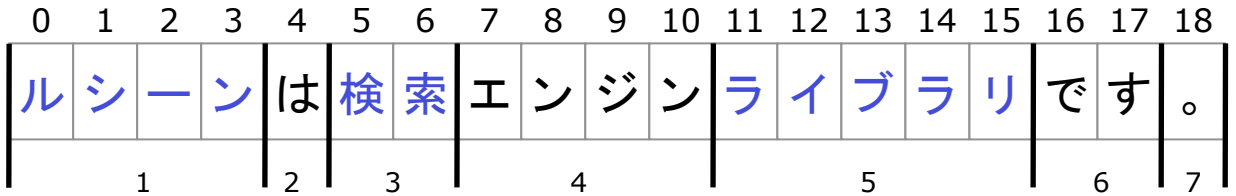
- ハイライトとは？

- トラブルその１
  「N-gramを使うとハイライトがおかしくなります」

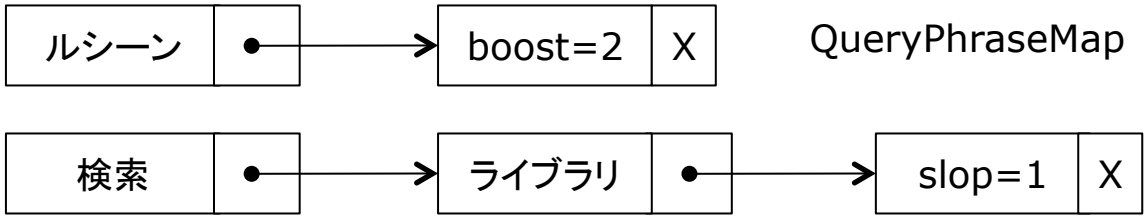
- トラブルその２
  「FVHを使うとハイライトがずれます」

FastVectorHighlighter

q=ルシーン^2 "検索 ライブラリ"~1




```
public class QueryPhraseMap {
    boolean terminal;
    int slop;
    float boost;
    Map<String, QueryPhraseMap> subMap;
}
```



FieldTermStack

ルシーン (0,4,1)
検索 (5,7,3)
ライブラリ (11,16,5)

FieldPhraseList

ルシーン [(0,4)] boost=2
検索 ライブラリ [(5,7),(11,16)] boost=1

FieldFragList

ルシーン [(0,4)]
検索 ライブラリ [(5,7),(11,16)]
totalBoost=3

Solrサブスクリプション・パッケージ

ロンウイットは、OSSのSolrを企業／団体様が安心してお使いいただけるよう、年間購読料形式の「サブスクリプション・パッケージ」としてご提供しています。
お客様のアプリケーションの目的に応じたエディションをご選択いただけます。

	BASIC	LANGUAGE	ENTERPRISE
Solr 1.4	√	√	√
バグフィックス	√	√	√
各種プラグイン	√	√	√
サポートサービス	√	√	√
言語判別		√	√
類義語辞書		√	√
固有表現抽出			√

Solrサブスクリプション・パッケージ

- 各種プラグイン　※提供予定の機能を含みます。また機能は順次追加されます。
  - 全角／半角／新旧漢字混じり、カタカナ語表記揺れ吸収検索
  - Java 6のNormalizerを使用した正規分解／互換分解／正規合成／互換合成処理による検索
  - 踊り字検索
  - ギャル文字検索
  - 顔文字検索
  - 1-gram最適化検索
  - 専門用語抽出
  - カタカナ⇔ひらがな⇔ローマ字検索
  - 日本語文章境界を認識するハイライター
  - 日本語サジェスチョン
  - Heritrixアーカイブファイルローダー

Solrサブスクリプション・パッケージ

- 言語判別
  - 世界およびインターネット上で使われている主要な下記25言語をサポート:

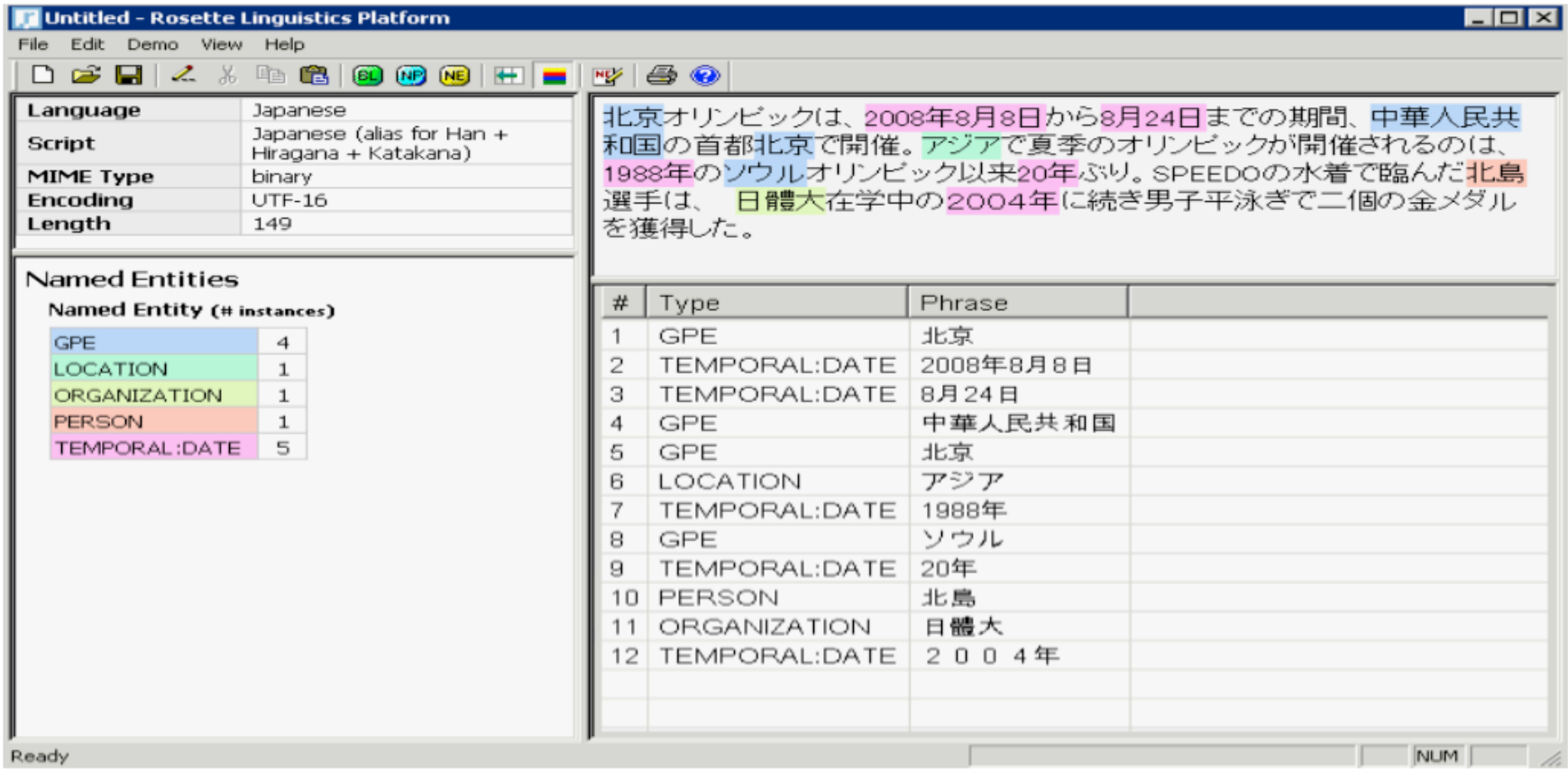
    アラビア語、ベンガル語、デンマーク語、ドイツ語、ギリシャ語、英語、スペイン語、ペルシャ語、フィンランド語、フランス語、ヒンディー語、ハンガリー語、アイスランド語、イタリア語、日本語、ジャワ語、韓国語、オランダ語、ノルウェー語、ポルトガル語、ロシア語、スウェーデン語、タイ語、トルコ語、中国語

- 類義語辞書
  - SynonymFilterで利用可能な分野別類義語辞書:

分野	およそのエントリ数
コンピューター	8,000
政治・経済	14,000
医療	3,000
科学	2,500
教育	3,000
芸能・スポーツ	2,000

Solrサブスクリプション・パッケージ

- 固有表現抽出 ※BASIS Technology社の固有表現抽出製品REXを用いています。
  - 人名／地名／国名／組織名などを文章から抽出
  - ベイズ理論に基づき、抽出すべき語の「文脈パターン」を事前学習



ロンウイットのSolr技術トレーニング

- 特徴
  - Lucene/Solrコミッター監修
  - 演習付き
  - PDFテキスト（持ち帰りOK）
  - 受講後1ヶ月のアフターサポート付き

- トレーニングコース　※2010年6月より順次ご提供できる予定です。
  - Solr 1.4 基礎
  - Solr 1.4 応用
  - Solr 1.4 運用（？）
  - Solr 1.4 プラグイン開発
  - Solr 1.4 DIH開発（？）

ご静聴ありがとうございました！