

Kuromoji

A new Japanese morphological analyzer

July 12th, 2011

Christian Moen
christian@atilika.com



Introducing Kuromoji

Kuromoji overview

- A Japanese morphological analyzer
 - Written in Java and is based on MeCab IPADIC, but has experimental support for UniDic
 - Open source with Apache License 2.0
- Why another morphological analyzer?
 - Could not find any actively maintained good pure Java analyzers when we started the project in 2010
 - Existing analyzers did not work great with search
 - Existing analyzers were not all that easy to use
- *These became our design goals for Kuromoji*



Key features

Basic morphological analysis

- Japanese text segmentation
- Part-of-speech tagging
- Lemmatization (by reduction)

東京都に住んでいます。

東京	名詞, 固有名詞, 地域, 一般, *, *, 東京, トウキョウ, トーキョー
都	名詞, 接尾, 地域, *, *, *, 都, ト, ト
に	助詞, 格助詞, 一般, *, *, *, に, ニ, ニ
住ん	動詞, 自立, *, *, 五段・マ行, 連用タ接続, 住む, スン, スン
で	助詞, 接続助詞, *, *, *, *, で, デ, デ
い	動詞, 非自立, *, *, 一段, 連用形, いる, イ, イ
ます	助動詞, *, *, *, 特殊・マス, 基本形, ます, マス, マス
。	記号, 句点, *, *, *, *, 。, 。, 。

Practical packaging

- Packaged as a self-contained jar (11MB)
 - Has no other 3rd party dependencies, including its statistical model and dictionaries
- Add kuromoji to classpath, and that's it!

```
% java -cp kuromoji-0.7.5.jar org.atilika.kuromoji.TokenizerRunner
Tokenizer ready.
すもももももものうち
すもも 名詞,一般,*,*,*,*,すもも,スモモ,スモモ
も 助詞,係助詞,*,*,*,*,も,モ,モ
もも 名詞,一般,*,*,*,*,もも,モモ,モモ
も 助詞,係助詞,*,*,*,*,も,モ,モ
もも 名詞,一般,*,*,*,*,もも,モモ,モモ
の 助詞,連体化,*,*,*,*,の,ノ,ノ
うち 名詞,非自立,副詞可能,*,*,*,うち,ウチ,ウチ
```

Might need to invoke java with -Dfile.encoding=UTF-8 depending on your platform

Designed for search

- Many analyzers do not segment compound nouns in a way generally useful for search
- Examples
 - 関西国際空港 and 日本経済新聞 are one token
 - Means we don't get a hit for 空港 and 新聞
- N-gramming is sometimes used address this
 - But has impacts on ranking (precision), index-size, performance, etc.

Designed for search (cont'd)

- Kuromoji has a segmentation mode that does additional splitting useful for search
 - 関西国際空港 ⇒ 関西 国際 空港
 - 日本経済新聞 ⇒ 日本 経済 新聞
- We can also unigram unknown words and unify morphological analysis and n-gramming

- デイジカメを買う ⇒

デ ィ ジ カ メ を 買う

unigrams

Apache Solr integration

- **Step 1:** Copy `kuromoji-0.7.6.jar` and `kuromoji-solr-0.5.3.jar` to the Solr lib directory
- **Step 2:** Define a type `text_ja` to `schema.xml` and define your Japanese fields using this type

```
<types>
  <fieldType name="text_ja" class="solr.TextField">
    <analyzer>
      <tokenizer class="org.atilika.kuromoji.solr.KuromojiTokenizerFactory"
        mode="search" user-dictionary="" />
    </analyzer>
  </fieldType>
</types>

<fields>
  <field name="title" type="text_ja" indexed="true" stored="true"/>
  <field name="body" type="text_ja" indexed="true" stored="true"/>
</fields>
```

- **Step 3:** Restart Solr, feed data and search

Easy to customize

- Own dictionaries can be used for ad hoc segmentation, i.e. to override default model
- File format is simple and there's no need to assign weights, etc. before using them
- Example custom dictionary:

```
# Custom segmentation and POS entry for long entries  
関西国際空港, 関西 国際 空港, カンサイ コクサイ クウコウ, カスタム名詞  
  
# Custom reading and POS former sumo wrestler Asashoryu  
朝青龍, 朝青龍, アサショウリュウ, カスタム人名
```

Mavenized

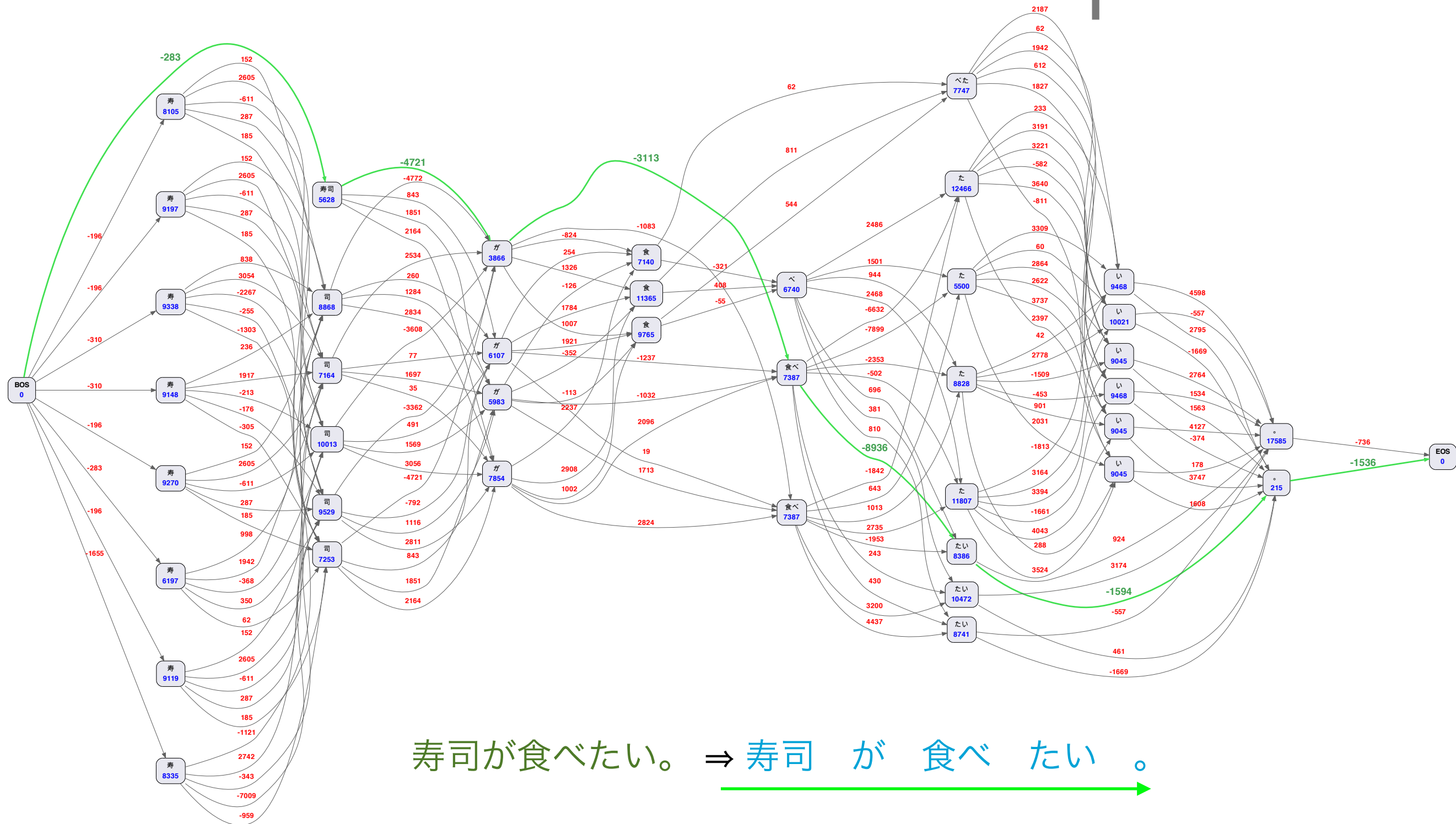
- Kuromoji is mavenized and is easy to use if you are using Maven (or Ant with Ivy)
- Add repository and dependency to pom.xml

```
<repositories>
  <repository>
    <id>ATILIKA Open Source</id>
    <url>http://atilika.org/nexus/content/repositories/atilika/</url>
  </repository>
</repositories>

<dependencies>
  <dependency>
    <groupId>org.atilika.kuromoji</groupId>
    <artifactId>kuromoji</artifactId>
    <version>0.7.5</version>
  </dependency>
</dependencies>
```

There's also a snapshots repository available on
<http://atilika.org/nexus/content/repositories/atilika-snapshots/>

Lattice and Viterbi path



寿司が食べたい。 ⇒ 寿司 が 食べ たい 。

Part-of-speech info is not shown in nodes

Kuromoji resources

- Homepage on
 - <http://atilika.org>
- Company homepage on
 - <http://atilika.com>
- Code is available on GitHub
 - <http://github.com/atilika/kuromoji> - Kuromoji
 - <http://github.com/atilika/kuromoji-solr> - Solr integration
 - <http://github.com/atilika/kuromoji-server> - Site demo