

# Package ‘openNLP’

October 26, 2019

**Encoding** UTF-8

**Version** 0.2-7

**Title** Apache OpenNLP Tools Interface

**Description** An interface to the Apache OpenNLP tools (version 1.5.3).

The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text written in Java.

It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution.

See <https://opennlp.apache.org/> for more information.

**Imports** NLP (>= 0.1-6.3), openNLPdata (>= 1.5.3-1), rJava (>= 0.6-3)

**Suggests** openNLPmodels.en

**Additional\_repositories** <https://datacube.wu.ac.at>

**SystemRequirements** Java (>= 5.0)

**License** GPL-3

**NeedsCompilation** no

**Author** Kurt Hornik [aut, cre] (<https://orcid.org/0000-0003-4198-9911>)

**Maintainer** Kurt Hornik <[Kurt.Hornik@R-project.org](mailto:Kurt.Hornik@R-project.org)>

**Repository** CRAN

**Date/Publication** 2019-10-26 15:52:51 UTC

## R topics documented:

Maxent_Chunk_Annotator . . . . .	2
Maxent_Entity_Annotator . . . . .	3
Maxent_POS_Tag_Annotator . . . . .	4
Maxent_Sent-Token_Annotator . . . . .	6
Maxent_Word-Token_Annotator . . . . .	7
Parse_Annotator . . . . .	8

<b>Index</b>	<b>10</b>
--------------	-----------

---

Maxent\_Chunk\_Annotator

*Apache OpenNLP based chunk annotators*

---

## Description

Generate an annotator which computes chunk annotations using the Apache OpenNLP Maxent chunker.

## Usage

```
Maxent_Chunk_Annotator(language = "en", probs = FALSE, model = NULL)
```

## Arguments

language	a character string giving the ISO-639 code of the language being processed by the annotator.
probs	a logical indicating whether the computed annotations should provide the token probabilities obtained from the Maxent model as their 'chunk_prob' feature.
model	a character string giving the path to the Maxent model file to be used, or NULL indicating to use a default model file for the given language (if available, see <b>Details</b> ).

## Details

See <http://opennlp.sourceforge.net/models-1.5/> for available model files. These can conveniently be made available to R by installing the respective **openNLPmodels**.*language* package from the repository at <https://datacube.wu.ac.at>.

## Value

An **Annotator** object giving the generated chunk annotator.

## See Also

<https://opennlp.apache.org> for more information about Apache OpenNLP.

## Examples

```
## Requires package 'openNLPmodels.en' from the repository at
## <https://datacube.wu.ac.at>.

require("NLP")
## Some text.
s <- paste(c("Pierre Vinken, 61 years old, will join the board as a ",
            "nonexecutive director Nov. 29.\n",
            "Mr. Vinken is chairman of Elsevier N.V., "),
```

```
        "the Dutch publishing group."),
      collapse = "")
s <- as.String(s)

## Chunking needs word token annotations with POS tags.
sent_token_annotator <- Maxent_Sent-Token_Annotator()
word_token_annotator <- Maxent_Word-Token_Annotator()
pos_tag_annotator <- Maxent_POS-Tag_Annotator()
a3 <- annotate(s,
  list(sent_token_annotator,
        word_token_annotator,
        pos_tag_annotator))

annotate(s, Maxent_Chunk_Annotator(), a3)
annotate(s, Maxent_Chunk_Annotator(probs = TRUE), a3)
```

---

Maxent\_Entity\_Annotator

*Apache OpenNLP based entity annotators*

---

## Description

Generate an annotator which computes entity annotations using the Apache OpenNLP Maxent name finder.

## Usage

```
Maxent_Entity_Annotator(language = "en", kind = "person", probs = FALSE,
                        model = NULL)
```

## Arguments

language	a character string giving the ISO-639 code of the language being processed by the annotator.
kind	a character string giving the ‘kind’ of entity to be annotated (person, date, ...).
probs	a logical indicating whether the computed annotations should provide the token probabilities obtained from the Maxent model as their ‘prob’ feature.
model	a character string giving the path to the Maxent model file to be used, or NULL indicating to use a default model file for the given language (if available, see <b>Details</b> ).

## Details

See <http://opennlp.sourceforge.net/models-1.5/> for available model files. These can conveniently be made available to R by installing the respective **openNLPmodels**.*language* package from the repository at <https://datacube.wu.ac.at>.

**Value**

An `Annotator` object giving the generated entity annotator.

**See Also**

<https://opennlp.apache.org> for more information about Apache OpenNLP.

**Examples**

```
## Requires package 'openNLPmodels.en' from the repository at
## <https://datacube.wu.ac.at>.

require("NLP")
## Some text.
s <- paste(c("Pierre Vinken, 61 years old, will join the board as a ",
            "nonexecutive director Nov. 29.\n",
            "Mr. Vinken is chairman of Elsevier N.V. ",
            "the Dutch publishing group."),
          collapse = "")
s <- as.String(s)

## Need sentence and word token annotations.
sent_token_annotator <- Maxent_Sent-Token_Annotator()
word_token_annotator <- Maxent_Word-Token_Annotator()
a2 <- annotate(s, list(sent_token_annotator, word_token_annotator))

## Entity recognition for persons.
entity_annotator <- Maxent_Entity_Annotator()
entity_annotator
annotate(s, entity_annotator, a2)
## Directly:
entity_annotator(s, a2)
## And slice ...
s[entity_annotator(s, a2)]
## Variant with sentence probabilities as features.
annotate(s, Maxent_Entity_Annotator(probs = TRUE), a2)
```

---

Maxent\_POS\_Tag\_Annotator

*Apache OpenNLP based POS tag annotators*

---

**Description**

Generate an annotator which computes POS tag annotations using the Apache OpenNLP Maxent Part of Speech tagger.

**Usage**

```
Maxent_POS_Tag_Annotator(language = "en", probs = FALSE, model = NULL)
```

**Arguments**

language	a character string giving the ISO-639 code of the language being processed by the annotator.
probs	a logical indicating whether the computed annotations should provide the token probabilities obtained from the Maxent model as their 'POS_prob' feature.
model	a character string giving the path to the Maxent model file to be used, or NULL indicating to use a default model file for the given language (if available, see <b>Details</b> ).

**Details**

See <http://opennlp.sourceforge.net/models-1.5/> for available model files. For languages other than English, these can conveniently be made available to R by installing the respective **openNLPmodels.language** package from the repository at <https://datacube.wu.ac.at>. For English, no additional installation is required.

**Value**

An **Annotator** object giving the generated POS tag annotator.

**See Also**

<https://opennlp.apache.org> for more information about Apache OpenNLP.

**Examples**

```
require("NLP")
## Some text.
s <- paste(c("Pierre Vinken, 61 years old, will join the board as a ",
            "nonexecutive director Nov. 29.\n",
            "Mr. Vinken is chairman of Elsevier N.V., ",
            "the Dutch publishing group."),
          collapse = "")
s <- as.String(s)

## Need sentence and word token annotations.
sent_token_annotator <- Maxent_Sent-Token_Annotator()
word_token_annotator <- Maxent_Word-Token_Annotator()
a2 <- annotate(s, list(sent_token_annotator, word_token_annotator))

pos_tag_annotator <- Maxent_POS_Tag_Annotator()
pos_tag_annotator
a3 <- annotate(s, pos_tag_annotator, a2)
a3
## Variant with POS tag probabilities as (additional) features.
head(annotate(s, Maxent_POS_Tag_Annotator(probs = TRUE), a2))
```

```

## Determine the distribution of POS tags for word tokens.
a3w <- subset(a3, type == "word")
tags <- sapply(a3w$features, `[[`, "POS")
tags
table(tags)
## Extract token/POS pairs (all of them): easy.
sprintf("%s/%s", s[a3w], tags)

## Extract pairs of word tokens and POS tags for second sentence:
a3ws2 <- annotations_in_spans(subset(a3, type == "word"),
                             subset(a3, type == "sentence")[2L])[1L]
sprintf("%s/%s", s[a3ws2], sapply(a3ws2$features, `[[`, "POS"))

```

---

Maxent\_Sent-Token\_Annotator

*Apache OpenNLP based sentence token annotators*

---

## Description

Generate an annotator which computes sentence annotations using the Apache OpenNLP Maxent sentence detector.

## Usage

```
Maxent_Sent-Token_Annotator(language = "en", probs = FALSE, model = NULL)
```

## Arguments

language	a character string giving the ISO-639 code of the language being processed by the annotator.
probs	a logical indicating whether the computed annotations should provide the token probabilities obtained from the Maxent model as their 'prob' feature.
model	a character string giving the path to the Maxent model file to be used, or NULL indicating to use a default model file for the given language (if available, see <b>Details</b> ).

## Details

See <http://opennlp.sourceforge.net/models-1.5/> for available model files. For languages other than English, these can conveniently be made available to R by installing the respective **openNLPmodels.language** package from the repository at <https://datacube.wu.ac.at>. For English, no additional installation is required.

## Value

An **Annotator** object giving the generated sentence token annotator.

**See Also**

<https://opennlp.apache.org> for more information about Apache OpenNLP.

**Examples**

```
require("NLP")
## Some text.
s <- paste(c("Pierre Vinken, 61 years old, will join the board as a ",
            "nonexecutive director Nov. 29.\n",
            "Mr. Vinken is chairman of Elsevier N.V., ",
            "the Dutch publishing group."),
          collapse = "")
s <- as.String(s)

sent_token_annotator <- Maxent_Sent-Token_Annotator()
sent_token_annotator
a1 <- annotate(s, sent_token_annotator)
a1
## Extract sentences.
s[a1]
## Variant with sentence probabilities as features.
annotate(s, Maxent_Sent-Token_Annotator(probs = TRUE))
```

---

Maxent\_Word-Token\_Annotator

*Apache OpenNLP based word token annotators*

---

**Description**

Generate an annotator which computes word token annotations using the Apache OpenNLP Maxent tokenizer.

**Usage**

```
Maxent_Word-Token_Annotator(language = "en", probs = FALSE, model = NULL)
```

**Arguments**

language	a character string giving the ISO-639 code of the language being processed by the annotator.
probs	a logical indicating whether the computed annotations should provide the token probabilities obtained from the Maxent model as their 'prob' feature.
model	a character string giving the path to the Maxent model file to be used, or NULL indicating to use a default model file for the given language (if available, see <b>Details</b> ).

**Details**

See <http://opennlp.sourceforge.net/models-1.5/> for available model files. For languages other than English, these can conveniently be made available to R by installing the respective **openNLPmodels.language** package from the repository at <https://datacube.wu.ac.at>. For English, no additional installation is required.

**Value**

An **Annotator** object giving the generated word token annotator.

**See Also**

<https://opennlp.apache.org> for more information about Apache OpenNLP.

**Examples**

```
require("NLP")
## Some text.
s <- paste(c("Pierre Vinken, 61 years old, will join the board as a ",
            "nonexecutive director Nov. 29.\n",
            "Mr. Vinken is chairman of Elsevier N.V., ",
            "the Dutch publishing group."),
          collapse = "")
s <- as.String(s)

## Need sentence token annotations.
sent_token_annotator <- Maxent_Sent-Token_Annotator()
a1 <- annotate(s, sent_token_annotator)

word_token_annotator <- Maxent_Word-Token_Annotator()
word_token_annotator
a2 <- annotate(s, word_token_annotator, a1)
a2
## Variant with word token probabilities as features.
head(annotate(s, Maxent_Word-Token_Annotator(probs = TRUE), a1))

## Can also perform sentence and word token annotations in a pipeline:
a <- annotate(s, list(sent_token_annotator, word_token_annotator))
head(a)
```

---

Parse\_Annotator

*Apache OpenNLP based parse annotator*


---

**Description**

Generate an annotator which computes Penn Treebank parse annotations using the Apache OpenNLP chunking parser for English.



## Usage

```
Parse_Annotator()
```

## Details

Using the generated annotator requires installing package **openNLPmodels.en** from the repository at <https://datacube.wu.ac.at> (which provides the Maxent model file used by the parser).

## Value

An [Annotator](#) object giving the generated parse annotator.

## See Also

<https://opennlp.apache.org> for more information about Apache OpenNLP.

## Examples

```
## Requires package 'openNLPmodels.en' from the repository at
## <https://datacube.wu.ac.at>.

require("NLP")
## Some text.
s <- paste(c("Pierre Vinken, 61 years old, will join the board as a ",
            "nonexecutive director Nov. 29.\n",
            "Mr. Vinken is chairman of Elsevier N.V., ",
            "the Dutch publishing group."),
          collapse = "")
s <- as.String(s)

## Need sentence and word token annotations.
sent_token_annotator <- Maxent_Sent-Token_Annotator()
word_token_annotator <- Maxent_Word-Token_Annotator()
a2 <- annotate(s, list(sent_token_annotator, word_token_annotator))

parse_annotator <- Parse_Annotator()
## Compute the parse annotations only.
p <- parse_annotator(s, a2)
## Extract the formatted parse trees.
ptexts <- sapply(p$features, `[`, "parse")
ptexts
## Read into NLP Tree objects.
ptrees <- lapply(ptexts, Tree_parse)
ptrees
```

# Index

Annotator, [2](#), [4-6](#), [8](#), [9](#)

Maxent\_Chunk\_Annotator, [2](#)

Maxent\_Entity\_Annotator, [3](#)

Maxent\_POS\_Tag\_Annotator, [4](#)

Maxent\_Sent-Token\_Annotator, [6](#)

Maxent\_Word-Token\_Annotator, [7](#)

Parse\_Annotator, [8](#)