

Package: RcppMeCab (via r-universe)

August 13, 2024

Title 'Rcpp' Wrapper for 'MeCab' Library

Version 0.1.0

Maintainer Akiru Kato <paithiov909@gmail.com>

Description R package based on 'Rcpp' for 'MeCab': Yet Another Part-of-Speech and Morphological Analyzer. The purpose of this package is providing a seamless developing and analyzing environment for CJK texts. This package utilizes parallel programming for providing highly efficient text preprocessing 'posParallel()' function.

License GPL (>= 3)

BugReports <https://github.com/paithiov909/RcppMeCab/issues>

Depends R (>= 3.4.0)

Imports gibasa, stats

Suggests roxygen2, testthat (>= 3.0.0)

Config/testthat/edition 3

Encoding UTF-8

Language en-US

RoxygenNote 7.2.3

SystemRequirements MeCab

Repository <https://paithiov909.r-universe.dev>

RemoteUrl <https://github.com/paithiov909/RcppMeCab>

RemoteRef HEAD

RemoteSha b311dd4098f7af62131cec4d2bc1afd05b370b12

Contents

pos	2
posParallel	3
pos_parallel	4
RcppMeCab	6

Index	7
--------------	----------

pos	<i>part-of-speech tagger</i>
-----	------------------------------

Description

pos returns part-of-speech (POS) tagged morphemes of the sentence.

Usage

```
pos(
  sentence,
  join = TRUE,
  format = c("list", "data.frame"),
  sys_dic = "",
  user_dic = ""
)
```

Arguments

sentence	A character vector of any length. For analyzing multiple sentences, put them in one character vector.
join	A logical to decide the output format. The default value is TRUE. If FALSE, the function will return morphemes only, and tags put in the attribute. if 'format="data.frame"', then this will be ignored.
format	A data type for the result. The default value is "list". You can set this to "data.frame" to get a result as data frame format.
sys_dic	A location of system MeCab dictionary. The default value is "".
user_dic	A location of user-specific MeCab dictionary. The default value is "".

Details

This is a basic function for MeCab part-of-speech tagger. The function gets a character vector of any length and runs a loop inside C++ to provide faster processing.

You can add a user dictionary to 'user_dic'. It should be compiled by 'mecab-dict-index'. You can find an explanation about compiling a user dictionary in the <https://github.com/junhewk/RcppMeCab>.

You can also set a system dictionary especially if you are using multiple dictionaries (for example, using both IPA and Juman dictionary at the same time in Japanese) in 'sys_dic'. Using `options(mecabSysDic="#the path to your system dictionary")`, you can set your preferred system dictionary to the R terminal.

If you want to get a morpheme only, use 'join = FALSE' to put tag names on the attribute. Basically, the function will return a list of character vectors with (morpheme)/(tag) elements.

Value

A string vector of POS tagged morpheme will be returned in conjoined character vector form. Element names of the list are original phrases

Examples

```
## Not run:
sentence <- c("some UTF-8 texts")
pos(sentence)
pos(sentence, join = FALSE)
pos(sentence, format = "data.frame")
pos(sentence, user_dic = "~/user_dic.dic")
# System dictionary example: in case of using mecab-ipadic-NEologd
pos(sentence, sys_dic = "/usr/local/lib/mecab/dic/mecab-ipadic-neologd/")

## End(Not run)
```

posParallel	<i>parallel version of part-of-speech tagger</i>
-------------	--

Description

posParallel returns part-of-speech (POS) tagged morphemes of the sentence.

Usage

```
posParallel(
  sentence,
  join = TRUE,
  format = c("list", "data.frame"),
  sys_dic = "",
  user_dic = ""
)
```

Arguments

sentence	A character vector of any length. For analyzing multiple sentences, put them in one character vector.
join	A logical to decide the output format. The default value is TRUE. If FALSE, the function will return morphemes only, and tags put in the attribute. if 'format="data.frame"', then this will be ignored.
format	A data type for the result. The default value is "list". You can set this to "data.frame" to get a result as data frame format.
sys_dic	A location of system MeCab dictionary. The default value is "".
user_dic	A location of user-specific MeCab dictionary. The default value is "".

Details

This is a parallelized version of MeCab part-of-speech tagger. The function gets a character vector of any length and runs a loop inside C++ with Intel TBB to provide faster processing.

Parallelizing over a character vector is not supported by RcppParallel. Thus, this function makes duplicates of the input and the output. Therefore, if your data volume is large, use pos or divide the vector to several sub-vectors.

You can add a user dictionary to 'user_dic'. It should be compiled by 'mecab-dict-index'. You can find an explanation about compiling a user dictionary in the <https://github.com/junhewk/RcppMeCab>.

You can also set a system dictionary especially if you are using multiple dictionaries (for example, using both IPA and Juman dictionary at the same time in Japanese) in 'sys_dic'. Using options(mecabSysDic="#the path to your system dictionary"), you can set your preferred system dictionary to the R terminal.

If you want to get a morpheme only, use 'join = FALSE' to put tag names on the attribute. Basically, the function will return a list of character vectors with (morpheme)/(tag) elements.

Value

A string vector of POS tagged morpheme will be returned in conjoined character vector form. Element names of the list are original phrases

Examples

```
## Not run:
sentence <- c("some UTF-8 texts")
posParallel(sentence)
posParallel(sentence, join = FALSE)
posParallel(sentence, format = "data.frame")
posParallel(sentence, user_dic = "~/user_dic.dic")
# System dictionary example: in case of using mecab-ipadic-NEologd
pos(sentence, sys_dic = "/usr/local/lib/mecab/dic/mecab-ipadic-neologd/")

## End(Not run)
```

pos_parallel

Alias of 'posParallel'

Description

posParallel returns part-of-speech (POS) tagged morphemes of the sentence.

Usage

```
pos_parallel(
  sentence,
  join = TRUE,
  format = c("list", "data.frame"),
  sys_dic = "",
  user_dic = ""
)
```

Arguments

sentence	A character vector of any length. For analyzing multiple sentences, put them in one character vector.
join	A logical to decide the output format. The default value is TRUE. If FALSE, the function will return morphemes only, and tags put in the attribute. if 'format="data.frame"', then this will be ignored.
format	A data type for the result. The default value is "list". You can set this to "data.frame" to get a result as data frame format.
sys_dic	A location of system MeCab dictionary. The default value is "".
user_dic	A location of user-specific MeCab dictionary. The default value is "".

Details

This is a parallelized version of MeCab part-of-speech tagger. The function gets a character vector of any length and runs a loop inside C++ with Intel TBB to provide faster processing.

Parallelizing over a character vector is not supported by RcppParallel. Thus, this function makes duplicates of the input and the output. Therefore, if your data volume is large, use pos or divide the vector to several sub-vectors.

You can add a user dictionary to 'user_dic'. It should be compiled by 'mecab-dict-index'. You can find an explanation about compiling a user dictionary in the <https://github.com/junhewk/RcppMeCab>.

You can also set a system dictionary especially if you are using multiple dictionaries (for example, using both IPA and Juman dictionary at the same time in Japanese) in 'sys_dic'. Using `options(mecabSysDic="#the path to your system dictionary")`, you can set your preferred system dictionary to the R terminal.

If you want to get a morpheme only, use 'join = FALSE' to put tag names on the attribute. Basically, the function will return a list of character vectors with (morpheme)/(tag) elements.

Value

A string vector of POS tagged morpheme will be returned in conjoined character vector form. Element names of the list are original phrases

Examples

```
## Not run:
sentence <- c("some UTF-8 texts")
posParallel(sentence)
posParallel(sentence, join = FALSE)
posParallel(sentence, format = "data.frame")
posParallel(sentence, user_dic = "~/user_dic.dic")
# System dictionary example: in case of using mecab-ipadic-NEologd
pos(sentence, sys_dic = "/usr/local/lib/mecab/dic/mecab-ipadic-neologd/")

## End(Not run)
```

RcppMeCab

RcppMeCab: Rcpp Wrapper for MeCab Library

Description

R package based on 'Rcpp' for 'MeCab': Yet Another Part-of-Speech and Morphological Analyzer (<http://taku910.github.io/mecab/>). The purpose of this package is providing a seamless developing and analyzing environment for CJK texts. This package utilizes parallel programming for providing highly efficient text preprocessing `posParallel()` function. For installation, please refer to [README.md](#) file.

Details

This package utilizes 'MeCab' C API and 'Rcpp' codes.

Author(s)

Junhewk Kim

References

- [MeCab](#)
- [Rcpp: Seamless R and C++ Integration](#)
- [Eunjeon project](#)

See Also

Useful links:

- Report bugs at <https://github.com/paithiov909/RcppMeCab/issues>

Index

- * **Chinese**

- RcppMeCab, [6](#)

- * **Japanese**

- RcppMeCab, [6](#)

- * **Korean**

- RcppMeCab, [6](#)

- * **MeCab**

- RcppMeCab, [6](#)

- * **morpheme**

- RcppMeCab, [6](#)

- * **nlp**

- RcppMeCab, [6](#)

- * **part-of-speech**

- RcppMeCab, [6](#)

pos, [2](#)

pos_parallel, [4](#)

posParallel, [3](#)

RcppMeCab, [6](#)

RcppMeCab-package (RcppMeCab), [6](#)