Package ‘RcppMeCab’

March 24, 2026

Title 'rcpp' Wrapper for 'mecab' Library

Version 0.0.1.5

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.
For installation, please refer to README.md file.

Depends R (>= 3.4.0)

License GPL

Encoding UTF-8

BugReports https://github.com/junhewk/RcppMeCab/issues

RoxygenNote 7.3.3

Language en-US

LinkingTo Rcpp, RcppParallel, BH

Imports Rcpp, RcppParallel

Suggests testthat, spelling

SystemRequirements MeCab 0.996 or higher for Japanese (libmecab-dev (deb), mecab-devel (rpm)), mecab-ko 0.999 (https://github.com/Pusnow/mecab-ko-msvc) for Korean

NeedsCompilation yes

Author Junhewk Kim [aut, cre],
Taku Kudo [aut],
Akiru Kato [ctb],
Patrick Schratz [ctb]

Maintainer Junhewk Kim <junhewk.kim@gmail.com>

Repository CRAN

Date/Publication 2026-03-24 06:00:02 UTC

Contents

`dict_index` *Compile a MeCab user dictionary*

Description

`dict_index` compiles a user dictionary CSV file into a binary dictionary that can be used with `pos` and `posParallel`.

Usage

```
dict_index(
  dic_csv,
  out_dic,
  dic_dir,
  dic_charset = "utf-8",
  out_charset = "utf-8"
)
```

Arguments

`dic_csv`	Character scalar. Path to the user dictionary CSV file(s). Multiple CSV files can be provided as a character vector.
`out_dic`	Character scalar. Path for the output compiled dictionary file.
`dic_dir`	Character scalar. Path to the system dictionary directory. This is required so that MeCab can reference the system dictionary configuration during compilation.
`dic_charset`	Character scalar. Charset of the input CSV file. Default is `"utf-8"`.
`out_charset`	Character scalar. Charset of the output dictionary. Default is `"utf-8"`.

Details

This function wraps MeCab's `mecab-dict-index` internally, so you do not need the command-line tool installed separately.

Value

Invisible `TRUE` on success.

Examples

```
## Not run:
dict_index(
  dic_csv = "user_words.csv",
  out_dic = "user.dic",
  dic_dir = "/usr/local/lib/mecab/dic/ipadic"
)

# Then use the compiled dictionary:
pos("some text", user_dic = "user.dic")

## End(Not run)
```

`download_dic` *Download and install a MeCab dictionary*

Description

Downloads and installs a MeCab system dictionary for the specified language. Japanese and Chinese dictionaries are compiled from source using the built-in `mecab-dict-index`; Korean dictionaries are downloaded pre-compiled. No system-level MeCab installation is required.

Usage

```
download_dic(lang)
```

Arguments

`lang`	Character scalar. Language code: `"ja"` for Japanese (IPAdic), `"ko"` for Korean (mecab-ko-dic), or `"zh"` for Chinese (mecab-jieba).

Details

Dictionaries are stored in the user data directory (`tools::R_user_dir("RcppMeCab", "data")`).

Value

Invisible path to the installed dictionary directory.

Examples

```
## Not run:
download_dic("ja")
download_dic("ko")
download_dic("zh")
pos("some text", lang = "ja")

## End(Not run)
```

list_dic *List installed MeCab dictionaries*

Description

Shows all available MeCab dictionaries, including the bundled dictionary and any downloaded via `download_dic`.

Usage

```
list_dic()
```

Value

A data frame with columns `lang`, `name`, `path`, and `active`.

Examples

```
## Not run:
list_dic()

## End(Not run)
```

pos *part-of-speech tagger*

Description

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

Usage

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

Arguments

sentence A character vector of any length. For analyzing multiple sentences, put them in one character vector.

join A bool 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.

lang Optional language code ("ja", "ko", or "zh") to select a dictionary installed via download_dic. When specified, this overrides sys_dic.

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=), 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 or a list of POS tagged morpheme will be returned in conjoined character vector form.

Examples

```
## Not run:
sentence <- c(#some UTF-8 texts)
pos(sentence)
pos(sentence, join = FALSE)
pos(sentence, format = "data.frame")
pos(sentence, lang = "ja")
pos(sentence, lang = "ko")
pos(sentence, sys_dic = "/path/to/custom/dic")
pos(sentence, user_dic = "/path/to/user.dic")

## End(Not run)
```

posParallel *parallel version of part-of-speech tagger*

Description

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

Usage

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

Arguments

sentence	A character vector of any length. For analyzing multiple sentences, put them in one character vector.
join	A bool 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.
lang	Optional language code ("ja", "ko", or "zh") to select a dictionary installed via download_dic. When specified, this overrides sys_dic.
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=), 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 or a list of POS tagged morpheme will be returned in conjoined character vector form.

Examples

```
## Not run:
sentence <- c(#some UTF-8 texts)
posParallel(sentence)
posParallel(sentence, join = FALSE)
posParallel(sentence, format = "data.frame")
posParallel(sentence, lang = "ja")
posParallel(sentence, lang = "ko")
posParallel(sentence, sys_dic = "/path/to/custom/dic")
posParallel(sentence, user_dic = "/path/to/user.dic")

## 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 Taku Kudo

References

- MeCab
- Rcpp: Seamless R and C++ Integration
- Eunjeon project

See Also

Useful links:

- Report bugs at https://github.com/junhewk/RcppMeCab/issues

`set_dic` *Set the active MeCab dictionary by language*

Description

Sets the default system dictionary used by `pos` and `posParallel`. This is equivalent to calling `options(mecabSysDic = path)` but allows selection by language code.

Usage

```
set_dic(lang)
```

Arguments

`lang`	Character scalar. Language code (`"ja"`, `"ko"`, or `"zh"`) or `"bundled"` to use the dictionary bundled with the package.

Value

Invisible path to the activated dictionary directory.

Examples

```
## Not run:
set_dic("ja")
pos("some Japanese text")

set_dic("ko")
pos("some Korean text")

## End(Not run)
```

Index