Package: RcppMeCab (via r-universe)

October 29, 2026

Title 'rcpp' Wrapper for 'mecab' Library

Version 0.0.1.7

Description R package based on 'Rcpp' for 'MeCab': Yet Another Part-of-Speech and Morphological Analyzer. It provides install-time engine profiles and dictionaries for Japanese, Korean, and Mandarin Chinese text. Runtime dictionary selection does not change the installed engine. 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

Config/roxygen2/version 8.0.0

LinkingTo Rcpp, RcppParallel, BH

Imports Rcpp, RcppParallel

Suggests testthat, spelling

SystemRequirements MeCab 0.996 or higher for Japanese and Chinese (libmecab-dev (deb), mecab-devel (rpm)), mecab-jieba 0.1.1 for Mandarin Chinese (https://github.com/lindera/mecab-jieba), 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>

Config/pak/sysreqs libmecab-dev

Repository https://cran.r-universe.dev

Date/Publication 2026-07-13 11:32:58 UTC

RemoteUrl https://github.com/cran/RcppMeCab

RemoteRef HEAD

RemoteSha 770c04f69979c2f224f81a49e5bcc74cd99a67df

Contents

dict_index 2
dictionary_info 3
download_dic 4
list_dic 5
pos 5
posParallel 7
RcppMeCab 8
set_dic 9

Index **10**

`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)
```

`dictionary_info` *Inspect loaded MeCab dictionaries*

Description

Returns metadata reported by MeCab for the active system and user dictionaries. This is useful for diagnosing which dictionary a package or R session is actually using.

Usage

```
dictionary_info(sys_dic = "", user_dic = "")
```

Arguments

`sys_dic`	Character scalar. System dictionary directory. When empty, the `mecabSysDic` option and then MeCab's default are used.
`user_dic`	Character scalar. Optional compiled user dictionary.

Value

A data frame with one row per loaded dictionary and columns `filename`, `charset`, `type`, `size`, `left_size`, `right_size`, and `version`.

Examples

```
## Not run:
dictionary_info()

## 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.

Installing a dictionary does not change the MeCab engine linked into RcppMeCab. Japanese and Chinese dictionaries use standard MeCab. Supported Korean behavior requires the mecab-ko engine selected at package installation.

Usage

```
download_dic(lang)
```

Arguments

`lang`	Character scalar. Dictionary 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 dictionary code ("ja", "ko", or "zh") selecting a dictionary installed via download_dic. This does not switch the MeCab engine. When specified, it 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.

The lang argument selects a dictionary; it does not switch the MeCab engine chosen when RcppMeCab was installed. Japanese and Chinese dictionaries use standard MeCab. Supported Korean behavior, including mecab-ko whitespace handling, requires the mecab-ko engine. Dictionary feature layouts are language-specific, and the historical data-frame columns do not expose all mecab-jieba metadata.

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 = "zh")
```

```
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 dictionary code ("ja", "ko", or "zh") selecting a dictionary installed via download_dic. This does not switch the MeCab engine. When specified, it 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.

The `lang` argument selects a dictionary; it does not switch the MeCab engine chosen when RcppMeCab was installed. Japanese and Chinese dictionaries use standard MeCab. Supported Korean behavior, including mecab-ko whitespace handling, requires the mecab-ko engine. Dictionary feature layouts are language-specific, and the historical data-frame columns do not expose all mecab-jieba metadata.

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 = "zh")
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/). It provides install-time engine profiles and dictionaries for Japanese, Korean, and Mandarin Chinese text. Runtime dictionary selection does not change the installed engine. 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*

Description

Sets the default system dictionary used by `pos` and `posParallel`. This is equivalent to calling `options(mecabSysDic = path)` but allows selection by dictionary code. It does not switch the MeCab engine linked into RcppMeCab. Japanese and Chinese dictionaries use standard MeCab; supported Korean behavior requires the mecab-ko engine selected at package installation.

Usage

```
set_dic(lang)
```

Arguments

`lang`	Character scalar. Dictionary 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("zh")
pos("some Chinese text")

## End(Not run)
```

Index

∗ **Chinese**
RcppMeCab, 8
∗ **Japanese**
RcppMeCab, 8
∗ **Korean**
RcppMeCab, 8
∗ **MeCab**
RcppMeCab, 8
∗ **morpheme**
RcppMeCab, 8
∗ **nlp**
RcppMeCab, 8
∗ **part-of-speech**
RcppMeCab, 8

dict_index, 2
dictionary_info, 3
download_dic, 4, *5–7*

list_dic, 5

pos, 5, *9*
posParallel, 7, *9*

RcppMeCab, 8
RcppMeCab-package *(*RcppMeCab*)*, 8

set_dic, 9