Package ‘RcppMeCab’

July 4, 2018

Title 'rcpp' Wrapper for 'mecab' Library
Version 0.0.1.2
Author Junhewk Kim [aut, cre],
        Taku Kudo [aut]
Maintainer Junhewk Kim <junhewk.kim@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.
For installation, please refer to README.md file.
Depends R (>= 3.4.0)
License GPL
Encoding UTF-8
LazyData true
BugReports https://github.com/junhewk/RcppMeCab/issues
LinkingTo Rcpp, RcppParallel, BH
Imports Rcpp, RcppParallel
SystemRequirements MeCab 0.996 (or mecab-ko 0.9.2) or higher, GNU make
RoxygenNote 6.0.1
NeedsCompilation yes
Repository CRAN
Date/Publication 2018-07-04 16:40:03 UTC

R topics documented:

pos ......................................................... 2
posParallel ............................................... 3
RcppMeCab .............................................. 4
Index ..................................................... 5
**Description**

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

**Usage**

```r
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 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"`, 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](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 of POS tagged morpheme will be returned in conjoined character vector form. Element name of the list are original phrases.
posParallel

Examples

## Not run:

```r
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  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"),
              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.
- `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](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 of POS tagged morpheme will be returned in conjoined character vector form. Element name of the list are original phrases.

**Examples**

```r
## Not run:
sentence <- c(#some UTF-8 texts)
posParallel(sentence)
posParallel(sentence, join = FALSE)
posParallel(sentence, format = "data.frame")
posParallel(sentence, user_dic = "/usr/local/lib/mecab/dic/mecab-ipadic-neologd/")

# System dictionary example: in case of using mecab-ipadic-NEologd
pos(sentence, sys_dic = "/usr/local/lib/mecab/dic/mecab-ipadic-neologd/")
```

**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/](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
Index

*Topic Chinese
  RcppMeCab, 4

*Topic Japanese
  RcppMeCab, 4

*Topic Korean
  RcppMeCab, 4

*Topic MeCab
  RcppMeCab, 4

*Topic morpheme
  RcppMeCab, 4

*Topic nlp
  RcppMeCab, 4

*Topic part-of-speech
  RcppMeCab, 4

pos, 2
posParallel, 3

RcppMeCab, 4
RcppMeCab-package (RcppMeCab), 4