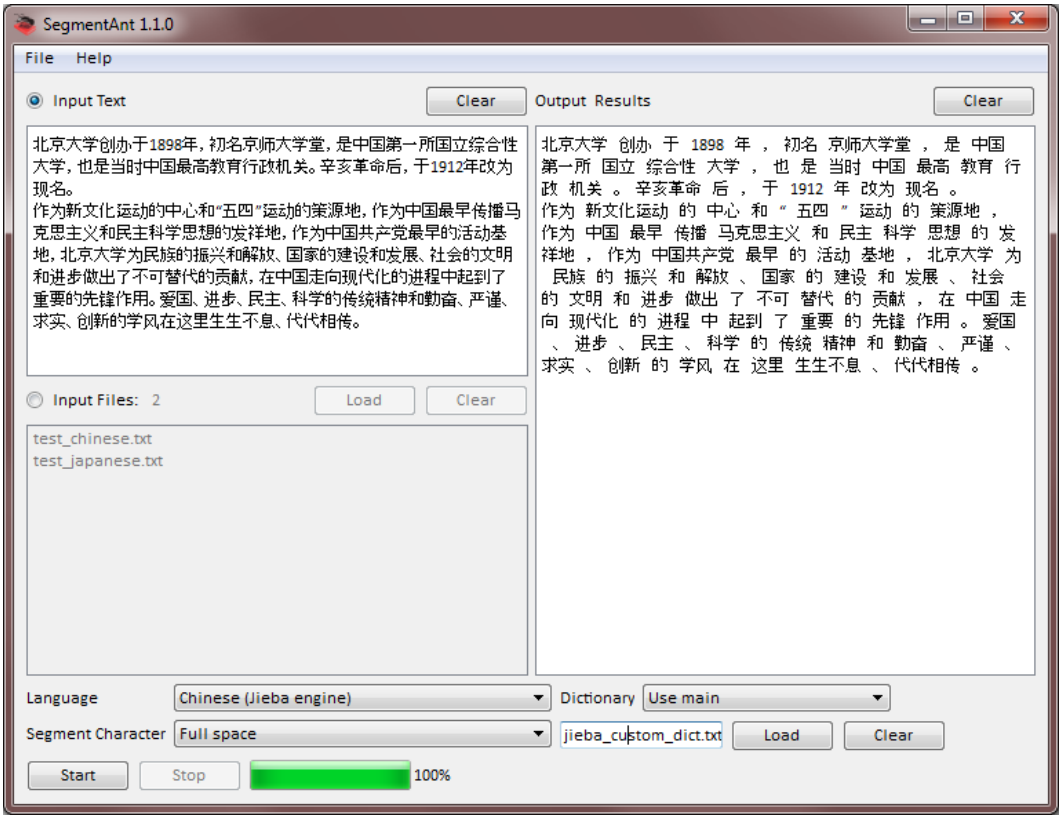
SegmentAnt (Windows)

Build 1.1.3 (Released October 27, 2017)

Laurence Anthony, Ph.D.
Center for English Language Education in Science and Engineering, School of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan.
Help file version: 001.



Introduction

SegmentAnt is a freeware Japanese and Chinese segmenting tool based on various tagging tools including the Jieba[1] and PyNLPIR (NLPIR/ICTLCAS)[2] engines for Chinese segmenting and POS tagging, the TinySegmenter[3] engine for Japanese segmenting, and the smallseg[4] engine for Chinese segmenting. *SegmentAnt* takes either an input text or an input list of text files (***UTF-8 encoded***) and splits the texts into 'tokens" separated by either half-width or full-width spaces. Depending on the embedded tool, *SegmentAnt* can also tag the texts with Parts-Of-Speech. *SegmentAnt* runs on any computer running Microsoft Windows (tested on Win 7), Macintosh OS X (tested on OS X 10.9 Mavericks), and Linux (tested on Linux Mint 17) computers. It is developed in Python and Qt using the *PyInstaller* compiler to generate executables for the different operating systems.

1 Available at: https://github.com/fxsjy/jieba

2 Available at http://pynlpir.readthedocs.org/en/latest/index.html.
The segmentation accuracy is reported to be 98.23% (http://ictclas.nlpir.org/nlpir/html/fenci-1.htmlt)

3 Developed by Taku Kudo <taku@chasen.org> and implemented in Python by Masato Hagiwara a athttp://lilyx.net/pages/tinysegmenterp.html>

4 Available at: https://code.google.com/p/smallseg/

Getting Started (No installation necessary)

Windows

On Windows systems, simply double click the *SegmentAnt* icon to launch the program.

Macintosh OS X

On Macintosh systems, simply double click the *SegmentAnt* zip file. The zip file will unzip the *SegmentAnt* application. Then, you can drag the *SegmentAnt* application to your application folder, your desktop, or anywhere else you like. Throw away the zip file when you are finished.

Linux

On Linux systems, set the permissions to run the executable, then double click the *SegmentAnt* icon to launch the program.

Segmenting Input Text

Step 1: Select the "Input Text" radiobox on the left of the main window.
Step 2: Select the language of the files using the combobox next to the "Language" label.
Step 3: Load a custom (user) dictionary if available for improved accuracy of segmenting/tagging. (See https://github.com/fxsjy/jieba for an explanation of how to use this feature with the Jieba engine).
Step 4: Choose either the "Half space" or "Full space" radiobox for the segment character of the output results.
Step 5: Click "Start" to begin the segmenting process.
Note 1: The segmenting process can be stopped at any time by clicking the "Stop" button.

Segmenting Input Files

Step 1: Select the "Input Files" radiobox on the left of the main window.
Step 2: Select the files you want to segment. You can do this in four ways:

a) Click on the File->Open File(s) menu option and select the files you want to segment;
b) Click on the File->Open Dir menu option and select a directory of files you want to segment;
c) Click on the "Load" button next to the "Input Files" label and select the files you want to segment;
d) Drag and drop files directly onto the *SegmentAnt* application.

Note 1: The number of selected files is shown next to the "Input Files" label.
Note 2: If you click on the File->Close Files menu option or click the "Close" button next to the "Input Files" label, the input files will removed from the list.

Step 3: Select the language of the files using the combobox next to the "Language" label.
Step 4: Load a custom (user) dictionary if available for improved accuracy of segmenting/tagging. (See https://github.com/fxsjy/jieba for an explanation of how to use this feature with the Jieba engine).
Step 5: Choose either the "Half space" or "Full space" radiobox for the segment character of the output results.
Step 6: Click "Start" to begin the segmenting process.

Note 1: Segmented versions of the original files will be saved in a new folder named "segmented" in the same folder as the original file.
Note 2: The segmenting process can be stopped at any time by clicking the "Stop" button.

Additional Features

The output display can be selected, copied, and pasted as is standard on the operating system:

Windows/Linux:	CTRL-A ⇨ Select All	CTRL-C ⇨ Copy	CTRL-V ⇨ Paste
Macintosh:	CMD-A ⇨ Select All	CMD -C ⇨ Copy	CMD -V ⇨ Paste

NOTES

Comments/Suggestions/Bug Fixes

All new editions and bug fixes are listed in the revision history below. However, if you find a bug in the program, or have any suggestions for improving the program, please let me know and I will try to address the issues in a future version.

This software is available as 'freeware' according to the license below. It is important for my funding to hear about any successes that people have with the software. Therefore, if you find the software useful, please send me an e-mail briefly describing how it is being used.

CITING/REFERENCING *SegmentAnt*

Use the following method to cite/reference *SegmentAnt* according to the APA style guide:

Anthony, L. (YEAR OF RELEASE). *SegmentAnt* (Version VERSION NUMBER) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

For example if you download *SegmentAnt 1.1.0*, which was released in 2014, you would cite/reference it as follows:
Anthony, L. (2015). *SegmentAnt* (Version 1.1.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http:// www.laurenceanthony.net/

Note that the APA instructions are not entirely clear about citing software, and it is debatable whether or not the "Available from ..." statement is needed. See here for more details:
http://owl.english.purdue.edu/owl/resource/560/10/

LICENSE for SegmentAnt

SegmentAnt 1.0 and any minor updates issued by AntLab Solutions (collectively 'the Software')

TERMS GOVERNING THE USE OF THE SOFTWARE

The Software is protected by copyright and must not be used, displayed, modified, adapted, distributed, transmitted, transferred or published or otherwise reproduced in any form by any means other than strictly in accordance with the terms set out below. By installing the Software, you agree to be bound by the terms of the license. This SegmentAnt License (“License”) is made between AntLab Solutions, Tokyo, Japan as licensor, and you, as licensee, as of the date of your use of the Software. The Software is in use on a computer when it is loaded into the RAM or installed into the permanent memory of that computer, e.g., a hard disk or other storage device.

1. License Material

These terms govern your use of the Software but not including subsequent versions (e.g. SegmentAnt 2.0').

2. License Grant

AntLab Solutions grants to you a personal non-exclusive non-transferable license ('the License') to use the Software in the following specific contexts.

a) Non-Commercial (Freeware) Use:

You may use the software for non-profit purposes on more than one computer or on a network so long as you are the sole user of the Software. (A “network” is any combination of two or more computers that are electronically linked and capable of sharing the use of a single software program.) You are not permitted to sell,

lease, distribute, transfer, sublicense, or otherwise dispose of the Software, in whole or in part, for any form of actual or potential commercial gain or consideration.

b) Commercial Evaluation (Trial) Use:
You may evaluate (trial) the software for commercial purposes for a period of no more than fourteen (14) days from the date of download on more than one computer or on a network so long as you are the sole user of the Software.

c) Commercial Use
When you pay the commercial license fee established by AntLab Solutions, you may use the software for non-profit or commercial purposes on more than one computer or on a network so long as you are the sole user of the Software. (A “network” is any combination of two or more computers that are electronically linked and capable of sharing the use of a single software program.) You will obtain a separate license for each additional user of the Software (whether or not such users are connected on a network). You are not permitted to sell, lease, distribute, transfer, sublicense, or otherwise dispose of the Software, in whole or in part, for any form of actual or potential commercial gain or consideration.

3. Termination
You may terminate this License at any time by uninstalling the Software and deleting it. The License will also terminate if you breach any of the terms of the License.

4. Proprietary Rights
The Software is licensed, not sold, to you. AntLab Solutions reserves all rights not expressly granted to you. Ownership of the Software and its associated proprietary rights, including but not limited to patent and patent applications, are retained by AntLab Solutions. The Software is protected by the copyright laws of Japan and by international treaties. Therefore, you must comply with such laws and treaties in your use of the Software. You agree not to remove any of AntLab Solutions' copyright, trademarks, and other proprietary notices from the Software.

5. Distribution
Except as may be expressly allowed in Section 2, or as otherwise agreed to in a written agreement signed by both you and AntLab Solutions, you will not distribute the Software, either in whole or in part, in any form or medium.

6. Transfer and Use Restrictions
You may not sell, license, sub-license, lend, lease, rent, share, assign, transmit, telecommunicate, export, distribute or otherwise transfer the Software to others, except as expressly permitted in this License Agreement or in another agreement with AntLab Solutions. You may not modify, reverse engineer, decompile, decrypt, extract, or otherwise disassemble the Software.

7. Warranties
ANTLAB SOLUTIONS MAKES NO WARRANTIES WHATSOEVER REGARDING THE SOFTWARE AND IN PARTICULAR, DOES NOT WARRANT THAT THE SOFTWARE WILL FUNCTION IN ACCORDANCE WITH THE ACCOMPANYING DOCUMENTATION IN EVERY COMBINATION OF HARDWARE PLATFORM OR SOFTWARE ENVIRONMENT OR CONFIGURATION, OR BE COMPATIBLE WITH EVERY COMPUTER SYSTEM. IF THE SOFTWARE IS DEFECTIVE FOR ANY REASON, YOU WILL ASSUME THE ENTIRE COST OF ALL NECESSARY REPAIRS OR REPLACEMENTS.

8. Disclaimer
ANTLAB SOLUTIONS DOES NOT WARRANT THAT THE SOFTWARE OR SERVICE IS FREE FROM BUGS, DEFECTS, ERRORS OR OMISSIONS. THE SOFTWARE OR SERVICE IS PROVIDED ON AN “AS IS” BASIS AND ANTLAB

SOLUTIONS MAKES NO OTHER WARRANTIES OR CONDITIONS, EXPRESS OR IMPLIED, WITH RESPECT TO THE SOFTWARE INCLUDING WITHOUT LIMITATION THE IMPLIED WARRANTIES OR CONDITIONS OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

9. Limitation of Liability
ANTLAB SOLUTIONS WILL HAVE NO LIABILITY OR OBLIGATION FOR ANY DAMAGES OR REMEDIES, INCLUDING, WITHOUT LIMITATION, THE COST OF SUBSTITUTE GOODS, LOST DATA, LOST PROFITS, LOST REVENUES OR ANY OTHER DIRECT, INDIRECT, INCIDENTAL, SPECIAL, GENERAL, PUNITIVE OR CONSEQUENTIAL DAMAGES, ARISING OUT OF THIS LICENSE OR THE USE OR INABILITY TO USE THE SOFTWARE OR SERVICE. IN NO EVENT WILL ANTLAB SOLUTIONS'S TOTAL AGGREGATE LIABILITY (WHETHER IN CONTRACT (INCLUDING FUNDAMENTAL BREACH), WARRANTY, TORT (INCLUDING NEGLIGENCE), PRODUCT LIABILITY, INTELLECTUAL PROPERTY INFRINGEMENT OR OTHER LEGAL THEORY) WITH REGARD TO THE SOFTWARE AND/OR THIS LICENSE EXCEED THE LICENSE FEE PAID BY YOU TO ANTLAB SOLUTIONS. FURTHER, ANTLAB SOLUTIONS WILL NOT BE LIABLE FOR ANY DELAY OR FAILURE TO PERFORM ITS OBLIGATIONS UNDER THIS LICENSE AS A RESULT OF ANY CAUSES OR CONDITIONS BEYOND ANTLAB SOLUTIONS' REASONABLE CONTROL

10. Jurisdiction
These terms will be governed by Japanese law and the Japanese courts shall have jurisdiction.

KNOWN ISSUES

If a very large file is copied to the "Input Text" box and segmented, once the results are generated, copying the results might use up all the available system memory and cause the program to crash. To avoid this problem, large files should always be segmented using the "Input File" option.

REVISION HISTORY

1.1.3
This is a minor upgrade with the following bug fix
Bug fixes

1. An expiry of the license to the NLPIR/ITCLCAS engine caused it to break inside SegmentAnt. This has now been fixed.

1.1.2
This is a minor upgrade with the following bug fix
Bug fixes

2. An update to the NLPIR/ITCLCAS engine (again) caused it to break inside SegmentAnt. This has now been fixed..

1.1.1
This is a minor upgrade with the following bug fix
Bug fixes

3. An update to the NLPIR/ITCLCAS engine caused it to break inside SegmentAnt. This has now been fixed by using version 0.4.6 of the engine.

1.1.0
This is a minor upgrade with new features and various bug fixes
New features

- Revised this help file and added a more formal license for the software

- Added the Jieba and PyNLPIR (NLPIR/ICTLCAS) Chinese segmentation/tagging engines allowing for more accurate tagging of files as well as the ability to Part-of-Speech (POS) tag the files.
- Adapted the interface to allow for custom dictionaries to be added that replace or append the main dictionary used for segmenting/tagging.
- Changed the output so that segmented/tagged files are saved to a new "segmented" sub-folder in the same folder as the original files
- Cleaned up the code to allow further improvements to be added more easily

1.0.0
This is the first version of the program

Copyright 2017 Laurence Anthony. All rights reserved.