Manual for the

Tagging Optimiser

Version 1.0

Author:

Martin Weisser

September 2018

Contents

1 Introduction

The Tagging Optimiser is a tool designed to allow corpus users to enhance and improve the output of freeware taggers. It takes one or more PoS-tagged files as input, tries to correct some of the errors produced by the probabilistic tagging systems that tend to form the basis for these taggers, and diversifies the tag set used to allow for more fine-grained grammatical analyses. At the same time, it makes the tags in the tagset more readable by expanding the traditional minimalistic tags, thereby making them more mnemonic and easier to understand.

It has been developed for and tested with the output from 3 different taggers, the TreeTagger (Schmid 1994), The Stanford POS Tagger (Toutanova et al. 2003), and my own Simple PoS Tagger (Weisser 2014), including the version integrated into my Simple Corpus Tool (Weisser 2018a).

2 Overview of the Tool

The layout of the tool is quite simple. It consists of a workspace for handling input and output files, as well as an editor pane for viewing/editing in- or output files. Figure 1 shows an image of the Tagging Optimiser startup screen.

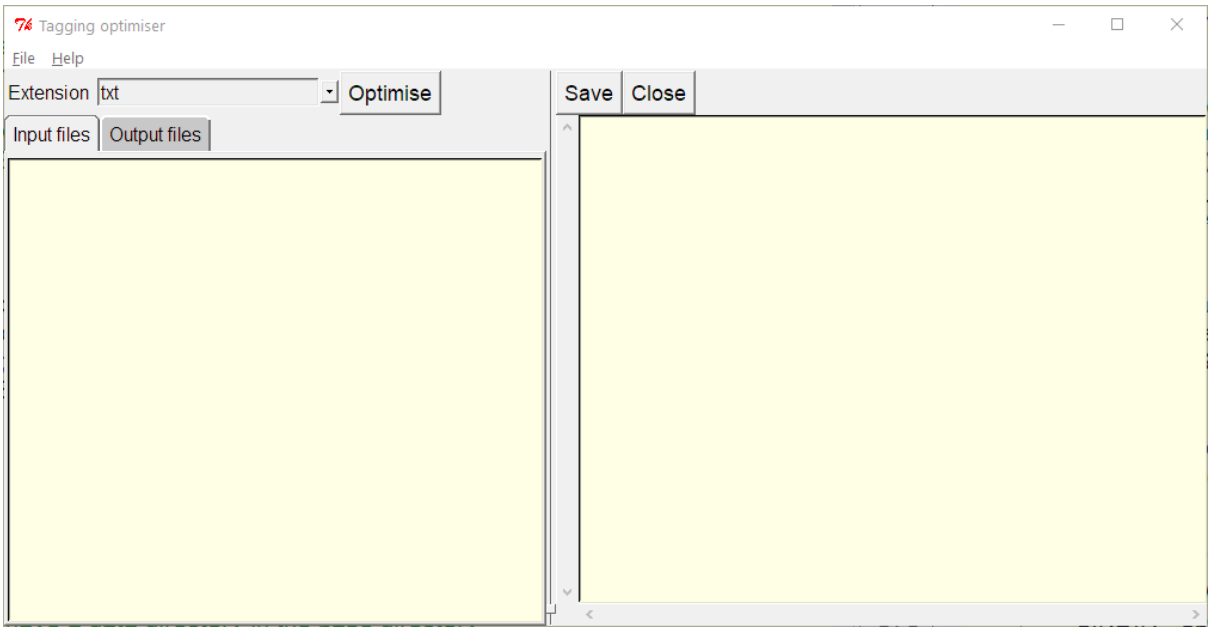

Figure 1 – The Tagging Optimiser Startup screen

The input/output workspace on the left contains 2 tabs, one where input files will be listed, and the other where any output files will be shown after optimisation. The editor pane on the right

consists of the editor window itself and 2 buttons, one for saving a file opened in the editor and the other for closing it.

3 Selecting Input Files

To select input files for optimising or editing, you first need to set an extension for the input files via the dropdown list above the input/output workspace. This option defaults to ‘txt’, but ‘xml’ can also be chosen, or user-defined extensions typed into the selection box.

There are 2 options for selecting files, either by choosing a whole directory/folder, or individual single or multiple files. To achieve the former, either use ‘File → Select source → Directory’ or press ‘Ctrl + Alt + d’ or ‘F2’ on the keyboard, then navigate to the relevant folder and press enter. All files that fit the extension defined earlier will then be listed under the ‘Input files’ tab in the input/output workspace.

For individual files, select ‘File → Select source → File(s)’ or press ‘Ctrl + Alt + f’. After navigating to the appropriate folder, you can then either select a single file or multiple ones, using the appropriate multi-selection mechanisms of your operating system. As before, all selected files will be listed in the input workspace.

Files listed in either workspace can also be selected and removed by pressing ‘Del’ in order to exclude them from processing.

The default file location for data files in the Tagging Optimiser is the ‘data’ sub-folder of the program folder, but of course you can also select any folder on your system that contains your data to be optimised.

4 Optimising Tagged Files

Once one or more files have been loaded into the input workspace, you can optimise them by clicking the ‘Optimise’ button. Provided you have write-access to the folder your data originates from, a new sub-folder called ‘optimised’ will then automatically be created and the optimised files output there, as well as listed under the ‘Output files’ tab, which will also automatically be activated. Should the latter not occur, it’s most likely that you don’t have write-access and you need to move your files to a different folder that you can write to.

5 Viewing/Editing Optimised or Input Files

Files loaded into the input or output workspace can be viewed or edited by double-clicking the filename in the respective workspace tab. They will then open in the editor window on the right. Any changes made to the file can be saved either via the 'Save' button or by pressing 'Ctrl + s'. You can close a file by clicking on the 'Close' button or pressing 'Ctrl + w'.

If you want to edit a number of files in succession, there's no need to close any open file first, though, as double-clicking a new file will automatically close the file before opening the new one, also prompting you to save or discard any unsaved changes. The same check will also be performed if you close the program window.

To facilitate the editing process, the editor window also has a number of keyboard shortcut bindings. The standard copy (Ctrl + c), cut (Ctrl + x), and paste operations (Ctrl + v) are automatically implemented via the editor widget, but, in addition, you can also use 'Ctrl + f' to find expressions and 'F3' to find the next occurrence of the same expression (provided it is still highlighted).

6 The Optimised Tagset (Ver. 1.0)

The table below list the entries in the optimised tagset. The right-hand column (labelled Example/Word(s)) either contains examples drawn from sample files of the FLOB or FROWN corpora I've used for testing, or, if the particular tag has a closed set of representatives, the complete list.

Please note that the final tagset for version 1.0 is slightly different from the one presented in Weisser (2018b), especially with regard to the tagging of participles, which are now labelled with the 'suffixes' *presPart* instead of *ing* and *pastPart* instead of *ppart* to make them easier to identify and search for in concordances.

Label	Explanation	Example/Word(s)
Det~def	determiner, definite (underspecified for number)	*the*
Det~indef	determiner, indefinite	*a*, *an*
Det~demSing	determiner, demonstrative singular	*this*, *that*

Label	Explanation	Example/Word(s)
Det~demPl	determiner, demonstrative plural	*these*, *those*
Noun~Sing	noun, general singular	*ability*, *creature*, *meeting*
Noun~Pl	noun, general plural	*consequences*, *people*, *workshops*
Noun~propSing	noun, proper singular	*English*, *Wales*
Noun~propPl	noun, proper plural	*Proms*, *Republicans*
Noun~propDaySing	noun, proper name of weekday singular	*Friday*
Noun~propDayPl	noun, proper name of weekday plural	*Fridays*
Verb~BEbase	verb, BE base form	*be*
Verb~BEpresPart	verb, BE present participle	*being*
Verb~BE1Sing	verb, BE 1st person singular	*am*
Verb~BE3Sing	verb, BE 3rd person singular	*is*
Verb~BEnot3Sing	verb, BE not 3rd person singular (underspecified for number & person)	*are*
Verb~BESingpast	verb, BE singular past tense	*was*
Verb~BEPlpast	verb, BE plural past tense	*were*
Verb~BEen	verb, BE past participle	*been*
Verb~HAVEbase	verb, HAVE base form	*have*
Verb~HAVEpresPart	verb, HAVE present participle	*having*
Verb~HAVE3Sing	verb, HAVE 3rd person singular	*has*
Verb~HAVEnot3Sing	verb, HAVE not 3rd person singular	*have*
Verb~HAVEpast	verb, HAVE past tense	*had*

Label	Explanation	Example/Word(s)
Verb~HAVEPerf	verb, HAVE perfect tense	*have* + Verb~pastPart
Verb~HAVEPastPerf	verb, HAVE past perfect	*had* + Verb~pastPart
Verb~DObase	verb, DO base form	*do*
Verb~DOpresPart	verb, DO present participle	*doing*
Verb~DO3Sing	verb, DO 3rd person singular	*does*
Verb~DOnot3Sing	verb, DO (underspecified for number & person)	*do*
Verb~DOpast	verb, DO past tense	*did*
Verb~DOen	verb, DO past participle	*done*
Verb~MOD	verb, modal	*may*, *can*
Verb~base	verb, general base form	*bring*, *go*, *tell*
Verb~presPart	verb, general present participle	*improving*, *making*
Verb~3Sing	verb, general 3rd person singular	*asks*, *gives*
Verb~not3Sing	verb, general not 3rd person singular (underspecified for person & number)	*call*, *say*
Verb~past	verb, general past tense	*asked*, *gave*
Verb~pastPart	verb, general past participle	*(had) looked*, *(had been) placed*
Adj	adjective, base form	*bumpy*, *glamorous*
Adj~comp	adjective, comparative form	*greater*, *less*
Adj~sup	adjective, superlative form	*best*, *clearest*
Num~card	number, cardinal	*eight*, *1986*
Num~ord	number, ordinal	*second*, *15th*
Quant	quantifier	*all*, *few*, *many*

Label	Explanation	Example/Word(s)
Adv	adverb, base form	*already*, *slightly*
Adv~comp	adverb, comparative form	*earlier*, *more*
Adv~sup	adverb, superlative form	*earliest*, *most*
Neg	negation operator	*not*, *n't*
Pron~1SingSub	pronoun, 1st person singular subject	*I*
Pron~1SingObj	pronoun, 1st person singular object	*me*
Pron~1PlSub	pronoun, 1st person plural subject	*we*
Pron~1PlObj	pronoun, 1st person plural object	*us*
Pron~2	pronoun, 2nd person (underspecified for number)	*you*
Pron~3SingMSubj	pronoun, 3rd person singular male subject	*he*
Pron~3SingMObj	pronoun, 3rd person singular male object	*him*
Pron~3SingFSubj	pronoun, 3rd person singular female subject	*she*
Pron~3SingFObj	pronoun, 3rd person singular female object	*her*
Pron~3SingN	pronoun, 3rd person singular neuter	*it*
Pron~3PlSub	pronoun, 3rd person plural subject underspecified for gender	*they*
Pron~3PlObj	pronoun, 3rd person plural object	*them*
Pron~1SingRefl	pronoun, 1st person singular reflexive	*myself*

Label	Explanation	Example/Word(s)
Pron~1PlRefl	pronoun, 1st person plural reflexive	*ourselves*
Pron~2Refl	pronoun, 2nd person reflexive underspecified for number	*yourself*
Pron~2PlRefl	pronoun, 2nd person plural reflexive	*yourselves*
Pron~3SingMRefl	pronoun, 3rd person singular reflexive	*himself*
Pron~3SingFRefl	pronoun, 3rd person singular reflexive	*herself*
Pron~3SingNRefl	pronoun, 3rd person singular reflexive	*itself*
Pron~3PlRefl	pronoun, 3rd person plural reflexive	*themselves*
Pron~rel	pronoun, relative	*that*, *what*, *which*, *who*, *whom*
Pron~relPoss	pronoun, relative reflexive	*whose*
Poss	possessive marker	*', 's*
Prep	preposition, general	*about*, *by*, *of*
Prep~comp	preposition, comparative	*than*
Part	particle	*down*, *out*, *up*
Inf	infinitive marker	*to*
Exist	existential	*there*
Con~co	conjunction, coordinating	*and*, *&, but*, *or*
Con~sub	conjunction, subordinating	*also*, *although*, *because*, *since*
Qword	question word	*what*, *whatever*, *who*
Resp~yes	response, yes	*yes*, *yep*, *yeah, aye*
Resp~no	response, no	*no*, *nope*
DM	discourse marker	*ok*, *okay*, *well*

Label	Explanation	Example/Word(s)
Interj	interjection	*ah*, *aw*, *oh*
Punc~stop	punctuation mark, full stop	.
Punc~exclam	punctuation mark, exclamation mark	!
Punc~query	punctuation mark, question mark	?
Punc~comma	punctuation mark, comma	,
Punc~semi	punctuation mark, semi-colon	;
Punc~colon	punctuation mark, colon	:
Punc~lquot	punctuation mark, left quotation mark	‘, ', “, "
Punc~rquot	punctuation mark, right quotation mark	’, ', ”, "
Punc~lbr	punctuation mark, opening bracket	(
Punc~rbr	punctuation mark, closing bracket	)
Sym	symbol	$, £, €

7 References

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Toutanova, K., Klein, D., Manning, C., and Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*. pp. 252–259.

Weisser, M. (2014). The Simple PoS Tagger (Version 1.0) [Computer Software]. Downloadable from http://martinweisser.org/ling_soft.html#tagger.

Weisser, M. (2018a). The Simple Corpus Tool (Version 2.0) [Computer Software]. Downloadable from http://martinweisser.org/ling_soft.html#viewer.

Weisser, M. (2018b). Automatically Enhancing Tagging Accuracy and Readability for Common Freeware Taggers. In Y. Tono & H. Isahara (Eds.). *Proceedings of APCLC 2018*, Takamatsu, Japan. 502–505.