User Manual

Stem Tool - Application for Manual Stem Annotation

Key Words

Stem - basic word form, obtained by suffix elimination

Stemmer - program which automatically annotates every word of an input document.

Stem Tool - program which is used to manually annotate stems for an input document.

Purpose

The goal of the application is to manually annotate stem for every word in an input document. This is an important part of research since it provides human labeled data which will later be used to measure performance of the Stemmers.

Workflow

Input

The input of the application is a file which has a sentence in each line. Tokens in sentences are separated by whitespaces. The input file represents a corpus for annotation. If there are already some annotated words, the program will also load mapping file (check Workflow - Output).

Output

The output of the program are two files:

1) Mapping file - Contains mapping word - stem. Also, contains contexts in which every word appears, as a list of strings separated by vertical line. The file is in TSV (Tab Separated Value) format. Columns are word, stem and contexts.

Line of the mapping file: word\t stem \t $context_1|context_2|...|context_n$

2) Replacing file - Contains same content as the input file, with each word replaced by its stem.

GUI Components

Menu

File - Open

Loading input file containing corpus. Loading mapping file <input_filename>_word_stem_map.txt. If mapping file does not exist, only corpus is loaded.

File - Save

Saves result in files (check Workflow - Output). Names of the output files are:

1) Mapping file: <input_filename>_word_stem_map.txt

2) Replacing file: <Input_filename>_replace.txt

View - Language

Choosing language of the program.

Context

U ListView komponenti su ispisani konteksti trenutne reči. U svakoj liniji se nalazi po jedan kontekst. Kontekst se formira od 5 reči sa leve i 5 reči sa desne strane trenutne reči. Ovo ne važi u situacijama kada se pojavi kraj rečenice pre nego što se izbroji 5 reči, sa bilo koje strane. U tom slučaju kontekst ide do kraja rečenice.

Button Split

Used in situations when there are words with same form in the input document, but have different stems. Example from Serbian language:

1. Selo ima hiljadu stanovnika. Stem for the noun "*selo*" is "*sel*" (consider: od sela, ka selu, seoski)

2. Dete je selo na stolicu. Stem for the verb "*sesti*" is "*se*" (consider: sednem, sedneš, seo, sela, selo, sesti)

The correct stem for the word "*selo*" depends on the part of speech word belongs to.

In such situations, we have to split words with different contexts. Button Split does just that. Clicking the button, two new words appear, with suffixes _1 and _2. If the already split word gets split, suffix gets the next number in sequence. In order for this button to perform, conditions are:

1) Select at least one row from the context ListView.

2) Leave at least one element of the context ListView unselected.

The button is only activated if conditions are met.

Current Word

Currently chosen word. Each letter of the word has its own box. Clicking the box, word is split to stem and the unused part. Stem is colored in blue, rest in red. The change can be seen in TextView component (check Workflow - GUI components - Text View). Initially, stem of the each word is equal to the word itself.

Arrows

Moving stem left and right. Can be used as an alternative to the word boxes.

Text View

Textual view of the current word's stem. Text written here is the stem which will be annotated when clicked on Make Stem. The reason for this is that sometimes stem cannot be obtained by simple suffix trimming.

Example: Word *uzročni*, stem *uzrok*.

Button Remove Stem

If the current word is stemmed, its stem is removed and the word moves to empty list.

Button Make Stem

Current word, with its stem written in TextView moves to the list of stemmed words.

Button Merge Groups

Performs inverse operation to the Split button. Words previously split can be merged again. **During merging, stem gets reset, it is equal to the word form. Also, list of split words gets updated, so that it is always from 1 to n**. For example, if there is a list a_1, a_2, a_3, a_4, merging a_2 and a_4 will create new list a_1, a_2, a_3. **Please check the word numeration after merging.** In order for this button to perform, conditions are:

1) Select at least two elements from the list of empty words or stemmed words.

2) Selected words should have same base. *dog_1, dog_2* is a valid choice, while a_1, no_2 is not.

The button is only activated if conditions are met.

List of Empty Words

Words from the input document which are not annotated.

List of Annotated (Stemmed) Words

Words that have stems annotated. Word is on the left side, stem on the right. It is necessary to save result to files, so that it does not get deleted after the application is closed.

Keyboard Shortcuts

ENTER - Make stem.

Ctrl + S - Save the result.

Left arrow, right arrow - Move stem left or right.

Running

Command Line: stemTool.jar

User Support

If any bugs or errors are noticed, please send an email to dejan.golubovic.27@gmail.com