Task Assignment LPP-LC 2020

Description

Our goal is to create a multilingual corpus of text and speech data based on the story “Le Pétit Prince” by Antoine de Saint-Exupéry.

The first two tasks pertain to the written part of the corpus, i.e. text.
The text version of the corpus needs special cleaning and formatting. This will be your first task. Our studies involve a set of questions to be asked at the end of each chapter in the spoken version, which need to be translated into the respective language. This is the second task.

The spoken part of the corpus will be created artificially using state-of-the art Text-to-Speech Synthesis (TTS). The best way to achieve this for many languages is by using Google’s Text-to-Speech API[1] and its natural sounding WaveNet[2] synthesizer based on current deep neural networks. However, we still require human input to judge whether the output sounds natural – and to manually modify the input in case it sounds unnatural. This is where we need your help as a native speaker!
The input to the Google system can be modified using the Speech Synthesis Markup Language[3] (SSML) to further increase the naturalness of the speech output.
The third task consists of synthesizing each sentence of the first 6 chapters of “The Little Prince” using a Google WaveNet voice. The online demo version is the fastest and most convenient way of trying out and deciding which markup is necessary for each sentence.
The SSML input text that you create will then go into our pipeline to create the spoken part of the corpus.
While the goal of task 1 is to create clean text, the goal of task 3 is to create good quality synthesis.

In the following we provide the exact task descriptions for all 3 parts. Please read these instructions carefully. If you have any questions regarding the task instructions, contact languagecycles@cbs.mpg.de

You do not need to finish the tasks in a specific order. Please submit the results of your work to languagecycles@cbs.mpg.de. You may also submit the results of each task separately, if you are finished with one task before starting another.

Payment will be carried out by invoice. After you have completed the tasks, send an invoice to MPI CBS. The detailed information on this as well as the corresponding e-mail address is enclosed in the official task assignment e-mail.

1 https://cloud.google.com/text-to-speech/?hl=en
2 https://cloud.google.com/text-to-speech/docs/wavenet
3 https://www.w3.org/TR/speech-synthesis11/#S3

Task 1: Manual correction of text files

In this task, you will use the file containing the translation of the story “Le Pétit Prince”.

Your task:

- create a cleaned text file from the raw text file (the entire story)
- remove the information from the title page
- segment the text so that each sentence is on a separate line (except very short phrases and ellipses that are part of a dialogue)
- remove picture captions
- keep the chapter numbers (use normal digits and not e.g. roman numerals)
- correct errors resulting from formatting and conversion, e.g. inserted or missing characters, page numbering, etc.
- make sure quotation marks are consistent
- replace unusual characters and symbols with standard characters
- beware of encoding issues **(unicode)**

You will need the cleaned sentences as a starting point for Task 3.

Please note:
Since this text is a story, it contains several dialogues between the characters. This means that some sentences can be very short or consist of several quotations.
In this case, be guided by the typical sentence length in the story as well as the punctuation in the text to determine what constitutes one sentence.
For example, the end of a sentence is usually marked with “.” but sometimes also with “...”, especially if the text continues with a capitalized word.

Submission format:

- one sentence per line
- complete text in one text file (**.txt**), named `lpp_cleaned_<language>.txt`
- please separate the chapters by a blank line followed by the chapter number

Task 2: Translation of questions and answers

Your task ist to translate a set of questions and answers from English to your language. Please make sure that the questions fit the storyline in the translated version and that they can be answered using this version of the story. Formulating questions that make sense has priority over translating the questions so that they are as close to the original text as possible. We will also provide the English translation of the story (“The Little Prince”) for reference.

*Please also make sure that the **vocabulary** (terms and phrases) used in the questions and answers are consistent with that used in the translated version of the story!*

Submission format:

- Please enter the questions into an excel sheet with the English original in the first column and your translations in the second column

Task 3: Using Google Text-to-Speech with SSML

For the third task, you will use the online demo version of Google Cloud Text-to-Speech: https://cloud.google.com/text-to-speech/

- select **`ssml`** for the input text window
- select your language
- select voice type “WaveNet”
- select voice name (female voice; see below)
- keep the default options in the second row (audio device profile, speed and pitch)

Your task:

- First listen to the **female WaveNet** voices available for your language, and select the voice that sounds the most natural to your ears (indexed by A/B/C etc.)
- enter each sentence of **chapter 1 to chapter 6** individually into the text field and listen to the synthesized speech output (“speak it”-button)
- use the SSML tags to manipulate the prosody of the synthesis so that it sounds natural and appropriate *(see SSML user guide)*
- your **goal** is to create natural sounding speech output and **at the same time**, reducing the amount of necessary SSML tags (i.e. remove redundant tags)
- once you have found a good markup for a sentence, copy the entire input sentence including SSML tags into a text file

Submission format:

- use **one .txt file per sentence**, and put the text **in one line** (please make sure that the markup does not get damaged if you use line breaks for the demo!)
- the text files should be named: `lpp_<name>_1.txt`, `lpp_<name>_2.txt`, … `<name>` stands for the voice name you are using, e.g. `lpp_en-GB-wavenet-D_1.txt`
- please put all files into **one folder**

Using the Speech Synthesis Markup Language

Official SSML documentation:
https://www.w3.org/TR/speech-synthesis11/#S3

User guide for the SSML support in Google Text-to-Speech:
https://cloud.google.com/text-to-speech/docs/ssml

Please read this guide carefully.

Please note:
The markup tags are only an optional extension for modifying the synthesis. They do not provide full control over the output. Often, the tags can override or interfere with each other. Since the Google synthesizer is also sensitive to punctuation, it can make sense to either insert or delete commas, full stops etc. in the text to obtain the desired output ***(In this case, please do not insert or remove punctuation in the cleaned text file from task 2!)*.**

It is your task to find the best markup for sentences that require additional modification to sound natural. You can share and discuss SSML “hacks” with the other native speakers currently working on other languages for our project in our **forum**:
https://lpp-synthesis.forumieren.de

Sign up for this discussion board with your name and native language so that we can recognize you. We will not use this forum for any other purpose than this task.