University of Macau

CISC7021 – **Applied Natural Language Processing**

Assignment 1, 2022/2023

(Due date: *27 September 2022*)

Introduction

In this assignment, we will prepare n-gram language models and evaluate the test set's perplexity. We will learn how to create a language model using the language model toolkit SRILM [1] (Stolcke, 2002). The toolkit can be downloaded at: http://www.speech.sri.com/projects/srilm/download.html. Basic instructions on using the SRILM toolkit can be found on the website also.

Train and Test Data

The training and testing data for this assignment come from the News Commentary, which is created to be used for training the English language model. The training data consists of 300 thousand lines of text. While the testing set consists of around 90 thousand lines of text. The data corpora are from the official website of *Shared Task: Machine Translation of News.*[2] Both the training and testing data can be downloaded from UMMoodle.

Tasks

1. Build word-based language models, 1-*gram*, 2-*gram,* and 3-*gram*, for English text given the training data, and measure the perplexity on the training and testing set.
2. Build character-based language models, 1-*gram* to 6-*gram*, using the training data and measuring the perplexity of the training and test set.
3. Collect more monolingual data from the *First Conference on Machine Translation* (WMT16) and add them to the training data. Build language models and measure the perplexity.

Environment Setup

We require all the related (development) tools for course assignments and projects are Linux/Unix programs. You need to have a Linux platform for conducting experiments and system implementation. Using a virtual machine (i.e. WM Virtual Box - https://www.virtualbox.org/) to host a Linux system (i.e. Ubuntu - http://www.ubuntu.com/) will be a good choice. We strongly recommend this. Besides, you will use different toolkits for various (pre)processing tasks in the coursework. For example, you need a g++ compiler for compiling the SRILM toolkit in this assignment.

---

[1] http://www.speech.sri.com/projects/srilm/download.html

[2] http://www.statmt.org/wmt16/translation-task.html

In any way, there are documents for using the toolkit. If you are new to processing text on the Linux platform, there is a very good introduction given by Church (1994)[3] of using Unix commands for basic text processing.

Report

You need to submit a report of your work (2~3 pages). It should clearly present what is going on in your experiments, how you achieve them, and solve problems you encountered. You should include tables (or graphs) of the data (e.g. corpora statistics), evaluated perplexities, etc. of your models. I am particularly interested to see the conclusions you draw about the models you made and the data you collected, as well as the analysis of the obtained results. The report should follow the two-column format of the ACL proceeding.[4,5]

References

1. Kenneth Ward Church. 1994. Unix™ for Poets. *Notes of a course from the European Summer School on Language and Speech Communication, Corpus Based Methods*.
2. Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing, ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA, September 16-20, 2002*.

[3] http://www.cs.upc.edu/~padro/Unixforpoets.pdf
[4] Formation information: https://mirror.aclweb.org/acl2015/files/acl2015.pdf
[5] http://acl2015.org/files/acl2015.tex or http://acl2015.org/files/acl2015.dot