MSc Computational Linguistics Thesis Proposal: Generating Automatic Scores of L2 Reading Difficulty for Chinese Social Media Content

Elye Bliss, Dec, 2022
Advisor: Prof. Fei Xia

Problem description

Over the course of their learning progression, second language (L2) learners encounter a major obstacle to overcome when trying to move from intermediate to more advanced levels: switching their input of materials from language-teaching materials prepared for the classroom to native materials such as books, movies, the news and social media created by and for other native speakers. This is a daunting task, as native materials vary considerably in level of difficulty for L2 learners. On the other hand, such materials could be more interesting, as well as relevant to the learners. They might try a jump-in-the-deep-end approach and engage with all content, regardless of its level of difficulty. Alternatively, they could selectively choose materials based on the level of difficulty–maximizing "comprehensible input"–but at the cost of having to devote time, and sometimes money, into finding these level-appropriate materials in the first place.

Recently re-popularized in online forums of L2 learners of Chinese, the comprehensible input (CI) approach follows from Krashen's "input hypothesis", first laid out in the 1970s. The CI approach of language acquisition encourages students to maximize their consumption of content in that language, namely of content that is at–or slightly below–their current level. In this theory, concepts such as "extensive reading" are used to describe reading large amounts of materials where the student is familiar with most all of the lexicon (98% is a commonly-cited

goal), which contrasts with “intensive reading”, whereby the reading material is more challenging. Materials that are too challenging not only discourage students, but also result in less overall materials consumed, and might actually be counterproductive to the goals of language and grammar acquisition. In short, CI emphasizes volume-maximizing over intensity-maximizing.

The input hypothesis is a strong-formed theory. Krashen went so far as to argue that CI is the sole requirement for achieving high levels of fluency. In more modern online discussions, teachers and students tend to reject this strong-form of the theory, while still acknowledging the overall value of maximizing CI. CI-maximizing remains a focus of common language-learning strategies, but are also supplemented with other forms of language practice.

For L2 learners of Chinese, graded readers have become popular ways of finding level-appropriate texts, while online forums also frequently discuss the levels of difficulty and interest of popular native materials such as books, movies and TV shows. Some small services have popped up such as online tools that can scan the contents of an e-book and return its unique word and character count. Meanwhile, there are few available resources on how to engage with online materials such as major social media sites, apps and viral content, without “jumping in the deep end”. Trends on mainland Chinese social media sites such as Weibo and WeChat, for example, garner significant attention by media worldwide and academic research (see list of papers section below), but students will have a hard time navigating the sheer volume of posts and shares at varying levels of difficulty. These are all areas where NLP tools could assist students. In particular, NLP could be used to automatically estimate the level of reading difficulty of online content, allowing students to better filter content that is grade-appropriate.

List of papers

Papers on classifying texts by reading difficulty:

- *Balyan et al, 2020*: Applying Natural Language Processing and Hierarchical Machine Learning Approaches to Text Difficulty Classification
- *Vasselli, 2019*: Automatic Scaling of Text for Training Second Language Reading Comprehension
- *Reynolds, 2016*: Insights from Russian second language readability classification complexity-dependent training requirements, and feature evaluation of multiple categories
- *Vajjala and Meurers, 2012*: On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition

Papers exploring trends topics and using NLP approaches to analyzing Chinese social media sites:

- *Cheng 2022*: Automatic Text Summarization for Public Health WeChat Official Accounts Platform Base on Improved TextRank
- *Chen et al., 2022*: Public Opinion Dynamics in Cyberspace on Russia–Ukraine War A Case Analysis With Chinese Weibo
- *Yang et al., 2022*: Spatial evolution patterns of public panic on Chinese social networks amidst the COVID-19 pandemic
- *Meihua et al., 2022*: An Analysis on the Weibo Topic Detection Based on K-means Algorithm
- *Zhu et al., 2022*: Revealing Public Opinion towards the COVID-19 Vaccine with Weibo Data in China BertFDA-Based Model
- *Lin et al, 2022*: A model study on predicting new COVID-19 cases in China based on social and news media
- *Pang et al., 2022*: Tackling fake news in socially mediated public spheres A comparison of Weibo and WeChat
- *Hou, 2022*: Research on Public Opinion on Twitter of 2022 Beijing Winter Olympics
- *Yang et al, 2022*: Topic identification and sentiment trends in Weibo and WeChat content related to intellectual property in China
- *Gao et al., 2022*: Differences of Challenges of Working from Home (WFH) between Weibo and Twitter Users during COVID-19
- *Deng et al., 2021*: Who is leading China's family planning policy discourse in Weibo? A social media text mining analysis
- *Han et al, 2021*: Weibo users perception of the COVID-19 pandemic on Chinese social networking service (Weibo) sentiment analysis and fuzzy-c-means model
- *Chen et al., 2021*: A Novel Machine Learning Framework for Comparison of Viral COVID-19–Related Sina Weibo and Twitter Posts Workflow Development and Content Analysis

- *Gu and Jiang, 2021*: Prediction of Political Leanings of Chinese Speaking Twitter Users
- *Han and Sun, 2021*: Exploring public attention about green consumption on Sina Weibo Using text mining and deep learning
- *Zhang 2020*: A Turbulent Decade The Changes in Chinese Popular Attitudes toward Democracy
- *Ng and Peng, 2020*: Linguistic Fingerprints of Internet Censorship the Case of Sina Weibo
- *Hu et al., 2020*: Weibo-COV A Large-Scale COVID-19 Social Media Dataset from Weibo
- *Hu et al., 2020*: Chinese Social Media Suggest Decreased Vaccine Acceptance in China An Observational Study on Weibo Following the 2018 Changchun Changsheng Vaccine Incident
- *Yang et al., 2020*: Cross-platform comparison of framed topics in Twitter and Weibo machine learning approaches to social media text mining
- *Xu et al., 2020*: Understanding Online Public Sentiments A Machine Learning-Based Analysis of English and Chinese Twitter Discourse during the 2019 Chinese National Day
- *Hswen et al., 2020*: Association of “covid19” Versus “chinese virus” With Anti-Asian Sentiments on Twitter March 9–23, 2020
- *Zeng et al., 2020*: Contested Chinese Dreams of AI? Public discourse about Artificial intelligence on WeChat and People’s Daily Online
- *Huang and Wang, 2019*: Building a Network to “Tell China Stories Well” Chinese Diplomatic Communication Strategies on Twitter
- *Bolsover and Howard, 2019*: Chinese computational propaganda automation, algorithms and the manipulation of information about Chinese politics on Twitter and Weibo
- *Ma et al., 2019*: A time-series based aggregation scheme for topic detection in Weibo short texts
- *Su 2019*: Exploring the effect of Weibo opinion leaders on the dynamics of public opinion in China A revisit of the two-step flow of communication
- *Zhang et al., 2018*: Nationalism on Weibo - Towards a Multifaceted Understanding of Chinese Nationalism
- *You et al., 2018*: Design of Data Mining of WeChat Public PlatformBased on Python
- *King, Gary, Jennifer Pan and Margaret E. Roberts. 2017*: How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument

Approaches presented in papers

A first look at research related to this specific second-language acquisition (SLA) task returns a mix of older and newer articles with large gaps, suggesting that it has not received consistent attention by NLP researchers. There is some literature on the topic of automated approaches to classifying reading difficulty when searching for papers on English as a first or second language (ESL). For example, Balyan et al., (2020) demonstrated the success of using a hierarchical approach to difficulty classification, revealing more complex linguistic features predictive of difficulty than traditional metrics. Examples of traditional metrics can be found in research such as Vajjala and Meurers, (2012), who explore readability assessment in SLA settings using lexical and syntactic measures, as well as word and sentence length. A more directly-relevant example is Reynolds (2016), who explores readability of Russian as an L2 using ML methods. Among lexical, morphological, syntactic and discourse features, Reynolds finds morphological features contain the highest information gain for distinguishing beginner from intermediate materials in Russian, however this finding does little to help the task for isolating languages. Collins-Thompson and Callan, 2004 derive a measure for grade classification based on a multinomial naïve bayes model, which could serve as a useful baseline model for this thesis project.

There is also a decent number of papers attempting to analyze trends on Chinese social media sites. These are only tertiarily related to the task of this master's thesis, yet I've explored many of them in order to learn common approaches to data collection on popular sites such as Weibo or apps such as WeChat. Data collection methods include using the official APIs, accessing private APIs via proxy, web crawlers, and archival sites. These papers cover a variety of research questions, such as identifying topics and keywords most likely to be censored, popular opinions on global issues such as the war in Ukraine, the COVID-19 pandemic, vaccines, or

sentiment analysis over niche topics such as property rights. They typically begin with a large corpus of posts collected over a date range of significance and containing certain keywords. From there, topics will be further identified with an approach such as Latent Dirichlet Allocation (LDA) for topic identification, the results of which then might then be grouped into larger topics via manual coding by the authors or automatic methods. The results of this exercise are then usually evaluated qualitatively by the authors in light of the research question, leading to conclusions that can be meaningful, but are also very open to interpretation.

Proposed thesis approach

This thesis proposes to facilitate the viewing of articles and social media posts filtered by level of difficulty. It involves two parts: creating an effective classifier for determining the reading difficulty of Chinese texts, and then applying that model to unseen real-world data sources such as articles shared on the social media app WeChat.

For the first step, existing graded readers that rely on human evaluation can be used as ground truth for supervised learning. One such graded reader is the popular news-reading website and app, the Chairman's Bao, which continuously adds to its existing corpus (currently 8,677 articles), graded by seven levels: the six levels of the standardized Chinese proficiency test, the Hanyu Shuiping Kaoshi (HSK), and one level for everything considered more difficult thank HSK 6. A month subscription to the Chairman's Bao costs under $10, which gives full access to its corpus. Another graded reader service is the Mandarin Companion, which provides longer stories that fall into three levels of difficulty. Using various sources with difference text styles and lengths will help to prevent overfitting. Once I have gathered a large enough corpus, I will split it into 80-20% train-test sets, and use k-fold cross-validation on the training set for model fitting.

One could imagine numerous linguistically-informed features that we would expect to help classify documents according to difficulty level, in addition to the traditional metrics cited in related research above. Unique vocabulary size and the word frequencies of the vocabulary used undoubtedly play a role in reading difficulty, and is the feature most commonly referred to as a determinant of reading difficulty in the comprehensible input hypothesis. For the levels of the HSK, it is also common for students to study reference lists of vocabulary that they should be expected to know at each level, which somewhat correspond to word frequencies. In Chinese, there is the additional consideration of individual characters and their frequencies. For

example, names of individuals frequently use rarer characters than common unigram and bigram words. Speculatively, other features might include named entities in general, sentence length, document length, and so on. TF-IDF as used for classic document-classification tasks should be a suitable starting point for this project, upon which I will experiment with more complex and linguistic features. The classifier might also need to be modeled differently for the task of classifying longer documents such as articles, versus sentences and paragraphs as found on social media posts.

Once I have a model that performs reasonably well on the training data, the next step of this thesis project will be to use the model to return automatically-generated reading-difficulty labels on native articles found in social media, on topics which language learners may find more engaging. This will mostly just be a proof of concept for the purpose of this thesis, rather than a fully-developed interactive feature. Ideally, this feature would be able to return various types of media from numerous platforms. However, in recent years, it has become nearly impossible for foreigners outside of China to establish working Weibo accounts, or access official APIs of Weibo and WeChat. In experiments performed so far, I have found some initial success using crawlers to obtain the most recent posts of users on Weibo, as well as using the proxy service mitmproxy to search public accounts posts on WeChat.

Plan for evaluation

The performance of the reading-difficulty classifier will be evaluated on a 20% hold-out set of labeled texts, such as those found in the Chairman's Bao corpus. Since this will be a 7-class multinomial classification task, precision, recall and F1 measures will not be applicable. Instead, this project will measure performance in terms of accuracy of the predicted labels of the test set.

Some very basic baseline measures against which to test model performance will include random choice predictions and majority class, using the most common difficulty level found in the training data. I will also compare the performances of several different models such as random forest, naïve bayes, and support vector machines, and make educated guesses as to why some might outperform others.

For the second part of the project, I could conduct an online poll of L2 Chinese learners on chinese-forums.com, requesting learners to evaluate a set of articles and their expected HSK level of difficulty. Additionally, I would be interested in performing a qualitative topic analysis of the articles and how they differ by reading level, and contain a bonus section discussing the results.