Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Thesis author	Anna Kriukova
Thesis title	Measuring readability of technical texts
Submitted	2022
Program	Computer Science **Specialization** Computational Linguistics

Review author	Silvie Cinkova **Role** advisor
Position	Institute of Formal and Applied Linguistics

Review text:

The thesis consists of four chapters surrounded by independent Introduction, Conclusion and attachments. The goal was to build a system for a commercial educational platform (Hyperskill) to help the content managers manage the quality of the learning materials in terms of readability for their target audience. The company delivered a dump of their data (learning texts, exercises, and user information such as their progress, spent time, and possible comments on the learning materials).

The introduction is very clear in explaining the motivation, the mainstream of the relevant research fields, as well as the structure of the entire document. All chapters contain a summary along with an outlook to the next chapter, and how they are related. This helps the reader to recognize and follow the proverbial red thread throughout the entire document.

Chapter 1 defines the main concepts; that is, **readability** and **audience**, and summarizes the related readability studies from the early stages to the state of the art. I would like to stress that the classic readability formulas are still in use and widely implemented in word processors and various writing aids, and the early assumptions as well as readability tips for authors are still valid.

Although seemingly based on very crude operationalizations (syllable, and token-per-sentence counts), the early formulas are linguistically informed by subtle features (e.g. semantic focus on persons vs. inanimate items). These are, however, only implicit in the formulas, since they were too difficult to automatically extract (especially in the pre-computer era). They were therefore replaced by these crude counts, which had luckily turned out to strongly correlate with them. Hence, the detailed information on the historical background as well as using the classic formulas as a baseline is absolutely appropriate.

Chapter 2 describes the Hyperskill data set. The level of detail is just right to follow Anna's reasoning, with no unnecessary digressions. The Hyperskill-specific terms are well explained, along with the writing guidelines that ensure a common stylistic register for all texts. Also, Anna carefully flags details that are going to be particularly important later.

Chapter 3 explains the classical formulas used and presents their (absent) correlations with the user statistics, which were normalized whenever appropriate (e.g. wrt text length), as well as correlations among the very user statistics. It also describes Anna's inventive domain-specific adaptation of the default lexicon of basic vocabulary underlying the Dale-Chall formula. It certainly revealed even more clearly that the Dale-Chall formula would not be useful, as the domain adaptation has even reversed the correlation, but the initial intuition, as well as the explanation of the unexpected result, makes perfect sense.

Chapter 4 describes three different setups for machine-learning experiments:

1. a binary prediction whether a text was ”good” or ”bad” according to a double manual assessment of the theory texts and a rule-based pattern extraction.
2. prediction of the ratio of users who would complete the step to those who opened it
3. prediction of seconds to complete, normalized by 10-token chunks, aggregated as the median value across the 200 most recent completions.

Anna used a range of linguistic features summarizing syntactic as well as lexical properties, all implemented in the QuitaUp tool developed at the Institute of the Czech National Corpus. Furthermore, she deployed the probability of each text according to the GPT-2 language model, as well as the statistical and metadata features known from the previous chapters. Last but not least, she approximated the conceptual difficulty of each text by topic modeling (Latent Semantic Analysis).

Anna has evaluated all three approaches and proposed a general pipeline in line with the current quality criteria applied by the quality managers at Hyperskill.

Chapter 4 ends with a realistic outlook at possible future extensions.

The conclusion is clear and well-written. The attachments increase the reproducibility of Anna's research.

Evaluation Summary

The thesis has a very clear structure. It is also written in a very comprehensible and correct language, framed by a neat typesetting. All these features make it easy to navigate and a pleasure to read. The methodological approach is well motivated. The abundant and carefully cited resources attest the truly impressive range of concepts and methods Anna mastered to achieve her goal.

The research goal itself presents just the appropriate level of ambition: Anna restricted her readability evaluation to a specific pool of texts, creatively exploiting the features at hand. Since the resulting readability prediction model performs well and the text pool is constantly growing, Anna's research has proven its potential for immediate practical use. I believe that **this thesis fully complies with the highest standards expected at the faculty, and I warmly recommend it to the defense**.

Specific questions, comments, and suggestions

These are only minor comments!

Content

Chapter 2: I struggled with the concept of **step**, which gets important in Section 3.5. Apparently it is a unit within a **topic**, and it can belong to either of two classes: **theory** or **task**. This could have been said more explicitly on its first mention on Page 13.

Chapter 3, Section 3.7 (Summary): the first item of the list would have benefited of a reference to Table 3.2. From the text itself it was not clear to me which set of statistics was meant: were *statistical metrics of texts themselves* being contrasted with somehow aggregated user statistics, or was it correlations of the readability formulas among themselves? It took me several times back and forth reading to associate this part with the corresponding experiment.

Chapter 4, Section 4.5.2: A very nice interpretation of the LSA features: no inherently problematic texts for any of the ten major topics identified by the LSA.

Punctuation

p. 5: *but it still, remains a challenging and presently topical task* → omit the comma

Article use

p. 5: *Marchisio et al. (2019) train *** machine translation system* → insert a determiner (indefinite article, possessive pronoun)

p. 11: I focus only on theoretical parts of topics → **the** theoretical parts (theory and tasks were introduced before)

p. 22: The table 3.3 → Table 3.3

Vocabulary/formulation

p. 6: *clearness* → *clarity*

p. 15: what part of students → which proportion of students

p. 16: STD → standard deviation, avg → average (arithmetic mean): why unexplained abbreviations, when they are anyway only used just three times each (including the figure caption) and concentrate on a small text part?

p. 30: *... there is no estimation of readability for the texts I am dealing with* → no comprehension tests have been performed on the texts I am dealing with.

p. 30: *results of reading tests* → *results of reading comprehension tests*

Typography

p. 11: One of the main content units is a *topic* - as the other key terms **theory** and **tasks** are bold, **topic** should also be bold rather than italic.

p. 15: topi_completion_rate → topic_completion_rate?

p. 18: Flesch reading ease → Flesch Reading Ease

p. 22: O'hayre → O'Hayre

p. 24: mentioned in section 2.1 → mentioned in Section 2.1

p. 27: figure 3.6 → Figure 3.6; Flesch reading ease → Flesch Reading Ease (in several places)

p. 35: in section 2.2 → in Section 2.2.

Possible overuse of the Saxon genitive with inanimate nouns, particularly apparent in plural

p. 11: dedicated to increasing *texts'* comprehensibility.

p. 14: *Texts'* statistics (but *Texts description* one subsection earlier)

p. 14: Along with *topics'* texts, I got access to some statistical metrics that are tracked
on the site. All of the metrics can be influenced by the theory *step's* readability

p. 18: guidelines for *topics'* theories

p. 18: for the *platform's* texts

The Saxon (aka *possessive*) genitive is preferred with animate nouns or pronouns. It is generally compatible with some classes of inanimate nouns, such as geographical, locative, and temporal nouns, but otherwise their collocability is limited to expressing possession. In all the cases listed above the relation is a quality or, at best, appurtenance, rather than possession. (This language suggestion is based on *A Comprehensive Grammar of the English Language* by Quirk et al.) Suggestions: try and vary *of*-genitive (*prerequisites of the topics*), prepositional phrases (*prerequisites for/in the topics*?), and compounds with singular (*topic prerequisites*).

I recommend the thesis for defense.

I suggest to consider the thesis for the annual award.

Anna has created an original interdisciplinary work that is both well theoretically founded and robustly implemented to serve a real-life purpose. The topic is traditional and yet very current, and so is Anna's approach to it. Last but not least, by persuading a commercial company to share their full data with her as an intern, Anna has proven substantial soft skills, which also deserves appreciation.

2022-07-28

Signature: Silvie Cinková