Manipulating the Difficulty of C-Tests
– Supplementary Material –

Ji-Ung Lee and **Erik Schwan** and **Christian M. Meyer**
Ubiquitous Knowledge Processing (UKP) Lab and Research Training Group AIPHES
Computer Science Department, Technische Universität Darmstadt, Germany
https://www.ukp.tu-darmstadt.de

This document provides supplementary material for our ACL 2019 paper "Manipulating the Difficulty of C-Tests".

1 C-Test Difficulty Manipulation

Feature description for Δ_{inc} and Δ_{dec}. We provide an extended feature description for the subset of features used for our relative difficulty prediction models Δ_{inc} and Δ_{dec}. Features marked with * are also used by the absolute difficulty prediction model proposed by Beinborn (2016). For a gap $g = (i, \ell)$ in word w_i, we define:

- the predicted absolute gap difficulty $d(g)$ for the initial C-test created with DEF obtained from our reproduced difficulty prediction system, see line 3 of algorithm 2 (PS),
- the word length $|w_i|$ (WL*),
- the new gap size $\ell \pm 1$ after modification (GL*),
- the modified character $w_i[\ell]$ when increasing or decreasing the gap (CH),
- a binary indicator if the gap is after a *th* sound (RG*), and
- the logarithmic difference of alternative solutions (LD*) capturing the change in the degree of ambiguity when increasing or decreasing ℓ.

Feature ablation test. We conduct feature ablation tests to evaluate the impact of each feature on our relative difficulty prediction models Δ_{inc} and Δ_{dec}. Both models were evaluated on all gap size combinations for 120 random texts from the Brown corpus (Francis, 1965) with a three-fold cross-validation. Table 1 shows the performance increase for each model after including each feature. RMSE shows the deviation and ρ the correlation of our relative difficulty prediction compared to the absolute difficulty prediction. Although the increase in performance with RG is not substantial, we decided to include it as a meaningful feature which measures the impact for increasing or decreasing the gap size in words starting with *th*.

	Δ_{inc}		Δ_{dec}	
Feature	RMSE	ρ	RMSE	ρ
PS	.088	.521	.213	.271
+ WL	.072	.712	.183	.570
+ GL	.066	.771	.162	.687
+ CH	.069	.735	.157	.707
+ RG	.069	.736	.157	.707
+ LD	.061	.805	.131	.806

Table 1: Feature ablation test for Δ_{inc} and Δ_{dec} compared to the full difficulty prediction system

2 Neural Network Parameters

Although obtaining state-of-the-art results in many tasks, the deep neural networks we evaluated during our preliminary experiments did perform worse than the SVM. We performed parameter tuning with 100 randomly initialized configurations for both, MLP and BiLSTM. We tune the following parameters:

- Number of hidden layers $H_l \in [1, ..., 5]$
- Number of hidden units $H_l^u \in [50, ..., 200]$
- Dropout rate $D_x \in [0.1, ..., 0.5]$

We use Adam with Nesterov Momentum (Dozat, 2016) as our optimizer and keep the batch size at 5 for both models. All models are trained for 200 epochs with an early stopping after 10 epochs with no improvement of the loss. Figure 1 shows the resulting architectures of both models after tuning. Since our goal is to output regression values, we use a linear activation function in the output layer.

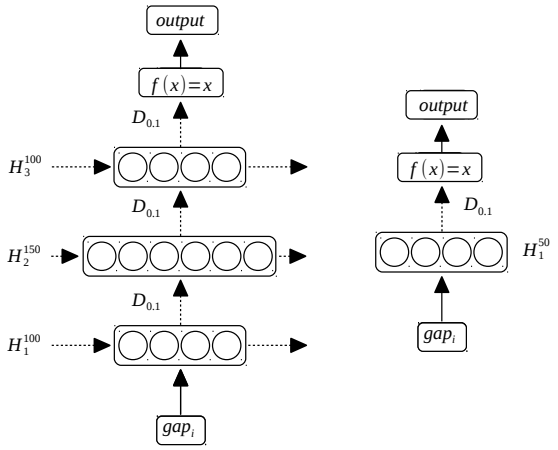

Figure 1: Final, tuned architectures of our BiLSTM (left) and MLP (right) models.

In preliminary experiments, we also tuned and evaluated BiLSTMs including soft attention, however, they performed even worse than the models without any attention. Analyzing the results of the best performing attention based model showed that it had a strong bias towards predicting the mean value of the whole training set. Furthermore, similar to the other neural models, it showed a low error on the training set (low bias) and a rather high error on the development set (high variance), indicating a lack of training data.

3 Evaluation of the Manipulation System

Results for additional corpora. Figure 2 and figure 3 show our results on the Gutenberg (Lahiri, 2014) and the Reuters (Lewis et al., 2004) corpora. As already discussed in the main paper, we observe very similar distributions for DEF, SEL, and SIZE across both corpora matching our descriptions for the Brown (Francis, 1965) corpus.

We further compute $\tau_{\max} - \tau_{\min}$ for SEL and SIZE for each text within a corpus and thus, measure the difficulty range both strategies are able to cover for a single text. As figure 4 shows, SEL achieves a larger difficulty range, whereas considerably more C-tests achieve higher difficulty levels when generated with SIZE. We again observe very similar distributions throughout the three corpora.

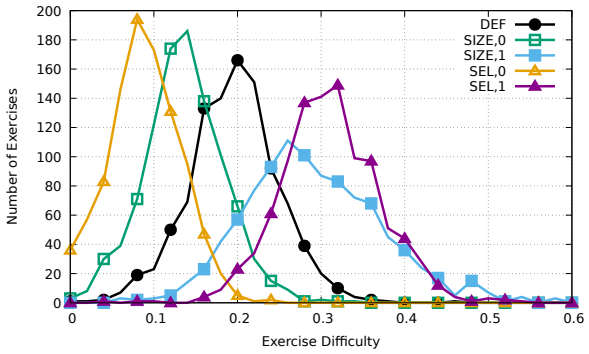

Figure 2: Difficulty distribution of exercises generated with DEF, SEL, and SIZE for extreme τ values on the Gutenberg corpus.

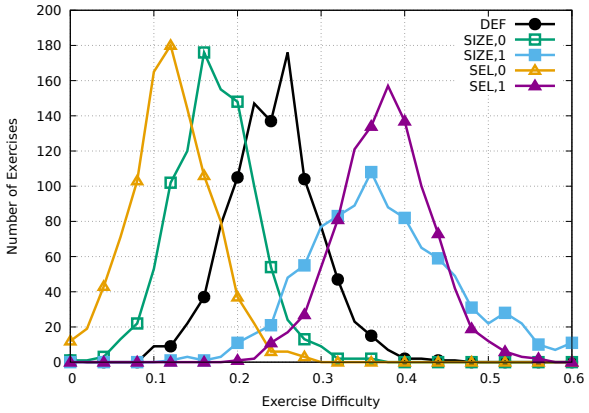

Figure 3: Difficulty distribution of exercises generated with DEF, SEL, and SIZE for extreme τ values on the Reuters corpus.

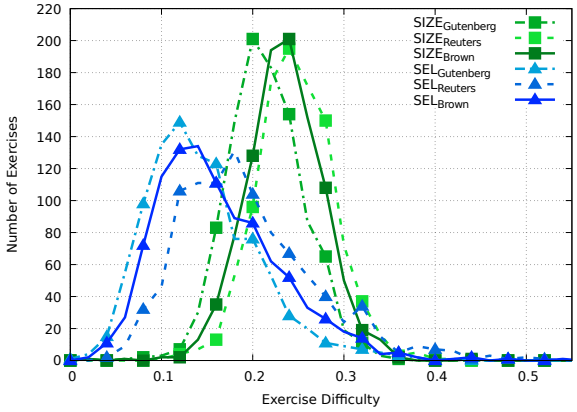

Figure 4: Error rate range ($\tau_{\max} - \tau_{\min}$) of exercises generated with SEL and SIZE for all three corpora.

4 User-based Evaluation

Questionnaire. At the begin of our study, our participants answered a questionnaire for a self-assessment of their English proficiency described in figure 5. We partitioned our questionnaire into three sections asking about 1) our participants' *English proficiency* (**Q1**, **Q2**), 2) their *learning habits and goals* (**Q4**), and 3) *other languages* they have been learning (**Q3**, **Q5**, **Q6**).

Q1: Please estimate your current language proficiency in English
A1: ○ *Beginner (A1)* ○ *Elementary (A2)* ○ *Intermediate (B1)* ○ *Upper Intermediate (B2)* ○ *Advanced (C1)* ○ *Proficiency (C2)*

Q2: I studied English for about ___ years.

Q3: Do you participate in any language learning courses (for example, at your university, evening school,...)? If yes, than which ones?
A3: ○ *Yes,* _____. ○ *No.*

Q4: How often do you practice English?
A4: ○ *Never* ○ *Monthly* ○ *Weekly* ○ *Daily*

Q5: What is your native language?
A5: _____

Q6: Have you tried learning other languages before? If yes, than which ones?
A6: ○ *Yes,* _____. ○ *No.*

Figure 5: Self-assessment questionnaire.

Answers. As described in the main paper, 17 participants are taking in language courses (**Q3**). Overall, 41 participants have tried to learn a second language (**Q6**). The exact answers can be found in the data we provide. Note, that not all participants provided the language which they attempted to learn since this was not mandatory. Figure 6–8 shows our participants' answers to **Q1**, **Q2**, and **Q4**. As can be seen, none of our participants consider themselves at the *Beginner (A1)* level. Furthermore, most of them are rather confident in their English proficiency and provide an estimate of either *Upper Intermediate (B2)* or *Advanced (C1)*.

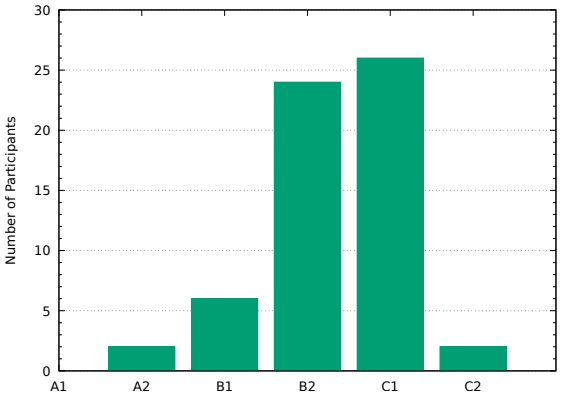

Figure 6: Our participants' CEFR level self-assessment

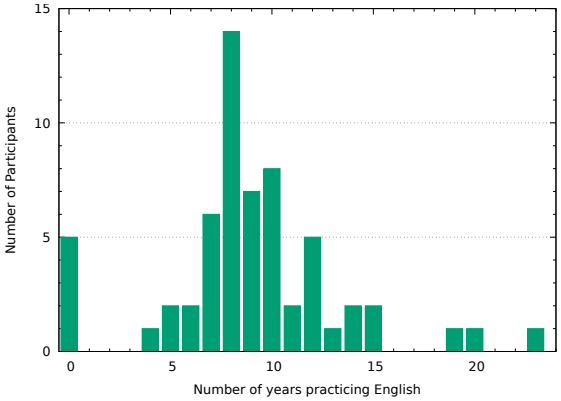

Figure 7: The number of years our participants have been practicing English

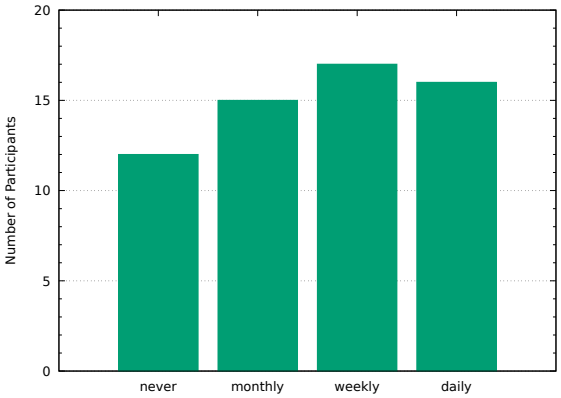

Figure 8: The frequency our participants have been practicing English

Readability index	T_1	T_2	T_3	T_4
Flesch reading ease	56.1	24.8	32	55.6
Gunning Fog	9.1	17.7	18.1	13.1
Flesch-Kincaid grade level	8.2	17.3	15.2	9.6
Coleman-Liau index	12	12	12	11
SMOG index	8.1	15.5	13.5	10.1
Automated readability index	7.9	17.4	15.5	9.7
Linsear Write formula	6.5	22.3	18.4	11.2

Table 2: Automated readability analysis of the four texts used for our C-tests. Scores are based on the online tool at http://www.readabilityformulas.com.

C-tests. Figure 9 shows the four texts T_1 to T_4 taken from the Brown corpus and the C-tests with the default gap scheme DEF we created from them for our user study. We have shortened each text to approximately 100 words and generated $n = 20$ gaps. In figure 10, we provide the results of our manipulation strategies SEL and SIZE with decreased ($\tau = 0.1$) and increased ($\tau = 0.5$) difficulty. Note that, we only show sentences that contain gaps; the beginning and end of each text is the same as in figure 9.

Table 2 reports readability scores for multiple common automated readability formulas. A Flesch reading ease score between 50–59 indicates *fairly difficult*, 30–49 *difficult*, and 0–29 *very difficult*. A Gunning Fog score of 9.1 indicates *fairly easy to read* and scores above 12 indicates *hard to read*. The remaining readability scores corresponding to grade levels.

The study of the St. Louis area's economic prospects prepared for the Construction Industry Joint Conference confirms and reinforces both the findings of the Metropolitan St. Louis Survey of 1957 and the easily observed picture of the Missouri-Illinois countryside. St. Louis si__ in t__ center o_ a relatively slow-growing a__ in so__ places stag____ mid-continent region . Slac_____ regional dem___ for St. Lo___ goods a__ services refl____ the reg___'s relative la__ of purch_____ power. N__ all St. Lo___ industries, o_ course, ha__ a market ar__ confined t_ the immediate neighborhood. But for those which do, the slow growth of the area has a retarding effect on the metropolitan core.

(a) C-test of T_1 with DEF gaps

Your invitation to write about Serge Prokofieff to honor his 70th Anniversary for the April issue of Sovietskaya Muzyka is accepted with pleasure, because I admire the music of Prokofieff; and with sober purpose, because the development of Prokofieff personifies, in many ways, the course of music in the Union of Soviet Socialist Republics. The Se___ Prokofieff wh__ we kn__ in t__ United Sta__ of Ame____ was g__, witty, merc_____, full o_ pranks a__ bonheur – a__ very cap____ as a profes_____ musician. Th___ qualities ende____ him t_ both t__ musicians a__ the social-economic ha___ monde wh___ supported the concert world of the post-World War 1, era. Prokofieff's outlook as a composer-pianist-conductor in America was, indeed, brilliant.

(b) C-test of T_2 with DEF gaps

The superb intellectual and spiritual vitality of William James was never more evident than in his letters. Here w__ a man wi__ an enor____ gift f__ living a_ well a_ thinking. T_ both per____ and id___ he bro____ the sa__ delighted inte____, the sa__ open-minded relish f__ what w__ unique i_ each, t__ same discrim_____ sensibility a__ quicksilver intell_____, the same gallantry of judgment. For this latest addition to the Great Letters Series, under the general editorship of Louis Kronenberger, Miss Hardwick has made a selection which admirably displays the variety of James's genius, not to mention the felicities of his style.

(c) C-test of T_3 with DEF gaps

Escalation unto death The nuclear war is already being fought, except that the bombs are not being dropped on enemy targets – not yet. It i_ being fou___, moreover, i_ fairly cl___ correspondence wi__ the predi_____ of t__ soothsayers o_ the th___ factories. Th_ predicted escal_____, and escal_____ is wh__ we a__ getting. T__biggest nuc____ device t__ United Sta___ has expl____ measured so__ 15 megatons, although our B-52s are said to be carrying two 20-megaton bombs apiece. Some time ago, however, Mr. Khrushchev decided that when bigger bombs were made, the Soviet Union would make them.

(d) C-test of T_4 with DEF gaps

Figure 9: Standard C-tests of our user study

. . . The Serg_ Prokofieff who_ we kne_ in t__ United State_ of Americ_ was ga_, witty, mercuria_, full o_ pranks an_ bonheur – an_ very capabl_ as a professiona_ musician. Thes_ qualities endeare_ him t_ both t__ musicians an_ the social-economic haut_ monde whic_ supported. . .	. . . The S____ Prokofieff wh__ we kn__ in t__ United S_____ of A_____ was ga_, witty, mercu____, full o_ pranks a__ bonheur – a__ very cap____ as a p__________ musician. T__ qualities end_____ him t_ both t__ musicians a__ the social-economic h____ monde wh___ supported. . .
(a) C-test of T_2 manipulated with SIZE for $\tau = 0.1$	(b) C-test of T_2 manipulated with SIZE for $\tau = 0.5$
. . . T__ Serge Proko_____ whom w_ kn__ i_ t__ Uni___ Sta___ o_ Ame___ w__ gay, witty, mercurial, fu__ o_ pranks and bonheur – a__ ve__ capable a__ a professional musician. These qualities endeared h__ t_ both t__ musicians a__ the social-economic haute monde which supported. . .	. . . The Se___ Prokofieff wh__ we kn__ in the United States of America was g__, wi___, merc_____, full of pra___ a__ bon____ – and very cap____ as a profes_____ musi____. Th___ qual_____ ende____ h__ to bo__ the musi____ and the social-economic ha___ mo___ which supported. . .
(c) C-test of T_2 manipulated with SEL for $\tau = 0.1$	(d) C-test of T_2 manipulated with SEL for $\tau = 0.5$
. . . Here wa_ a man wit_ an enormou_ gift fo_ living a_ well a_ thinking. T_ both person_ and idea_ he brough_ the sa__ delighted interes_, the sa__ open-minded relish fo_ what wa_ unique i_ each, t__ same discriminatin_ sensibility an_ quicksilver intelligenc_, the same gallantry of judgment. . .	. . . Here w__ a man w___ an e_____ gift f__ living a_ well a_ thinking. T_ both per____ and id___ he bro____ the s___ delighted inte____, the s___ open-minded relish f__ what w__ unique i_ each, t__ same d__________ sensibility a__ quicksilver i__________, the same gallantry of judgment. . .
(e) C-test of T_3 manipulated with SIZE for $\tau = 0.1$	(f) C-test of T_3 manipulated with SIZE for $\tau = 0.5$
. . . Here w__ a m__ wi__ a_ enormous gift f__ liv___ a_ we__ a_ thinking. T_ both persons and ideas h_ bro____ t__ sa__ delighted interest, t__ sa__ open-minded relish f__ what w__ unique i_ each, t__ same discriminating sensibility and quicksilver intelligence, the same gallantry of judgment. . .	. . . He__ was a m__ with an enor____ gi__ for living as well as thin____. T_ bo__ per____ a__ id___ he brought the same deli_____ inte____, the same open-minded rel___ for wh__ was uni___ in ea__, the same discrim_____ sensi_____ a__ quick_____ intelligence, the same gallantry of judgment. . .
(g) C-test of T_3 manipulated with SEL for $\tau = 0.1$	(h) C-test of T_3 manipulated with SEL for $\tau = 0.5$
. . . It i_ being fough_, moreover, i_ fairly clos_ correspondence wit_ the prediction_ of t__ soothsayers o_ the thin_ factories. The_ predicted escalatio_, and escalatio_ is wha_ we ar_ getting. T__ biggest nuclea_ device t__ United State_ has explode_ measured som_ 15 megatons. . .	. . . It i_ being fou___, moreover, i_ fairly c____ correspondence w___ the p__________ of t__ soothsayers o_ the th___ factories. T___ predicted es_________, and es_________ is wh__ we a__ getting. T__ biggest nu_____ device t__ United Sta___ has expl____ measured s___ 15 megatons. . .
(i) C-test of T_4 manipulated with SIZE for $\tau = 0.1$	(j) C-test of T_4 manipulated with SIZE for $\tau = 0.5$
. . . I_ i_ be___ fou___, moreover, i_ fairly close correspondence wi__ t__ predictions o_ t__ soothsayers o_ t__ think factories. They predicted escalation, a__ escalation i_ wh__ w_ a__ getting. T__ big____ nuclear device t__ Uni___ States has exploded measured some 15 megatons. . .	. . . It is being fought, more____, in fai___ cl___ corresp_____ with the predi_____ of the sooth_____ of the th___ fact_____. Th__ pred_____ escal_____, and escal_____ is what w_ are get_____. The big____ nuc____ dev___ the United States h__ expl____ meas____ some 15 megatons. . .
(k) C-test of T_4 manipulated with SEL for $\tau = 0.1$	(l) C-test of T_4 manipulated with SEL for $\tau = 0.5$

Figure 10: Manipulated C-tests of our user study

References

Lisa Marina Beinborn. 2016. *Predicting and manipulating the difficulty of text-completion exercises for language learning*. Ph.D. thesis, Technische Universität Darmstadt.

Timothy Dozat. 2016. Incorporating nesterov momentum into adam. In *ICLR Workshop*.

W. Nelson Francis. 1965. A standard corpus of edited present-day american english. *College English*, 26(4):267–273.

Shibamouli Lahiri. 2014. Complexity of Word Collocation Networks: A Preliminary Structural Analysis. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 96–105, Gothenburg, Sweden.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5(Apr):361–397.