

Regression or classification? Automated Essay Scoring for Norwegian

Stig Johan Berggren, Taraka Rama, Lilja Øvrelid

University of Oslo, Norway (`{stigbj — taraka.kasi}@gmail.com, liljao@ifi.uio.no`)



In short

First results for the task of Automated Essay Scoring for Norwegian learner language. We analyze a number of properties of this task experimentally and assess

- (i) the formulation of the task as either regression or classification,
- (ii) the use of various non-neural and neural machine learning architectures with various types of input representations, and
- (iii) applying multi-task learning for joint prediction of essay scoring and native language identification.

We find that a GRU-based attention model trained in a single-task setting performs best at the AES task.

The ASK corpus

- Norwegian learner essays from two different language tests which test proficiency at the B1 and B2 levels.

First language	AL test	IL test	Total
English	100	100	200
Polish	100	100	200
Russian	100	100	200
Somali	7	100	107
Spanish	100	100	200
German	100	100	200
Vietnamese	5	100	105
Subtotal (included languages)	512	700	1212
(Albanian)	24	100	124
(Bosnian-Croatian-Serbian)	100	100	200
(Dutch)	100	100	200
(Norwegian nynorsk)	21	11	32
(Norwegian bokmål)	79	89	168
Subtotal (excluded languages)	324	400	724
Total (all languages)	836	1100	1936

Models

- Linear models: Regression, Classification
- Neural: CNNs, RNNs, Attention
- Multi-tasking model: NLI as auxiliary task

AES: Linear

Model	All labels		Collapsed labels	
	Macro F ₁	Micro F ₁	Macro F ₁	Micro F ₁
Majority	0.040	0.163	0.127	0.341
LogReg BOW	0.199	0.317	0.384	0.626
LogReg Char	0.221	0.317	0.399	0.602
LogReg POS	0.190	0.301	0.312	0.569
LogReg Mix	0.213	0.341	0.337	0.577
SVC BOW	0.210	0.317	0.391	0.610
SVC Char	0.189	0.293	0.347	0.537
SVC POS	0.157	0.244	0.336	0.618
SVC Mix	0.215	0.350	0.319	0.585
SVR BOW	0.444	0.415	0.429	0.659
SVR Char	0.252	0.317	0.440	0.602
SVR POS	0.334	0.358	0.476	0.593
SVR Mix	0.312	0.350	0.441	0.659

AES: RNNs

Model	All labels		Collapsed labels	
	Macro F ₁	Micro F ₁	Macro F ₁	Micro F ₁
Random init, unidirectional GRU				
Mean	0.264	0.374	0.455	0.675
Max	0.219	0.325	0.487	0.683
Attn	0.434	0.431	0.806	0.805
+POS Mean	0.348	0.398	0.450	0.642
+POS Max	0.230	0.374	0.500	0.748
+POS Attn	0.434	0.423	0.718	0.813
Mix Mean	0.225	0.333	0.388	0.634
Mix Max	0.200	0.398	0.398	0.756
Mix Attn	0.302	0.455	0.509	0.780
Random init, BiGRU				
Mean	0.314	0.333	0.444	0.667
Max	0.160	0.325	0.460	0.691
Attn	0.459	0.447	0.805	0.805
+POS Mean	0.373	0.333	0.425	0.683
+POS Max	0.175	0.309	0.503	0.748
+POS Attn	0.460	0.447	0.687	0.821
Mix Mean	0.231	0.350	0.395	0.642
Mix Max	0.200	0.382	0.405	0.764
Mix Attn	0.275	0.455	0.617	0.707
Pre-trained, unidirectional GRU				
Mean	0.274	0.366	0.463	0.715
Max	0.185	0.350	0.401	0.756
Attn	0.414	0.431	0.678	0.797
+POS Mean	0.282	0.382	0.477	0.699
+POS Max	0.193	0.382	0.405	0.764
+POS Attn	0.409	0.423	0.746	0.789
Pre-trained, BiGRU				
Mean	0.266	0.390	0.435	0.707
Max	0.187	0.398	0.393	0.740
Attn	0.454	0.447	0.773	0.797
+POS Mean	0.281	0.382	0.480	0.724
+POS Max	0.183	0.341	0.397	0.748
+POS Attn	0.433	0.439	0.758	0.805

Attention visualization

Oppgave A | Din helsestatus har mye å si om hvordan du opplever ditt livskvalitet. Det er mange måter å ta vare på sin egen helse og UNK viser at mange mennesker lever lenger enn tidligere på grunn av ny kunnskap og utvikling innenfor medisin og teknologi. Spørsmålet er: hvordan kan vi ta vare på vår helse og er det greit å oppnå en høy alder? Det er flere måter at du kan ta vare på din egen helse. Først kan du passe på det du spiser. Alle vet nå at fett er usunt, men hvor mange leser UNK for å vite om UNK av maten de spiser? Å bli kjent med kunstige UNK er også viktig. Noen av disse midler er ikke farlig men, på den andre siden, er andre UNK. Annen kan du lære om de tingene du trenger for å bli sterk og sunt i kroppen. Mye er skrevet i det siste om, f.eks. vitaminer og god og skadelig kolesterol. Les i hvis du tar vare på kroppen din, skal du kanskje oppnå en høy alder. Men, er dette nødvendigvis en god ting? Jeg synes det har mye å si for samfunnet. For det første, å ha eldre mennesker som en del av samfunnet er bra for samfunnet når det gjelder livserfaring. Vi har eller kan, lære mye fra de eldre. På den andre siden, er det en stor belastning for samfunnet (på grunn av bestrøringen til helsevesen) å ha så mange eldre mennesker. Ofte trenger eldre mennesker mye omsorg senere i livet og mange opplever langvarig opphold i sykehus eller i UNK. For de eldre selv er

Essay by English native speaker

Oppgave A | I dag bor vi i en fantastisk tid. Man får masse muligheter. Vi har mye av det som besteforeldrene våre kunne bare drømme om. Det er helt utrolig å tenke på hva vi har oppnådd. Men samtidig har vi fått mange problemer. En av de problemene er vår egen helse. I dag har vi et godt utviklet UNK og mange forskjellige UNK. Men all dette hjelper neppe så mye hvis hvert enkeltmenneske ikke tenker på sin egen helse. Vi må huske at livet er en gave og en god helse hjelper veldig mye for å nyte livet. Derfor er det viktig å tenke på den måten vi lever på. Først og fremst bør man tenke på hva man spiser. Undersøkelser viser at folk har færre UNK i de landene hvor man spiser mye grønnsaker og frukt. Å ha UNK UNK til middag UNK for brus er også veldig sunt. Ikke minst viktig er det å ha litt mosjon og trim. I dag bruker man mye transport. Vi kjører veldig mye selv om vi ikke trenger det. Jeg var veldig overasket da jeg kom første gang til Norge og oppdaget at mennesker bruker bil for å komme. Mange har sånn jobb som vi ikke trenger å sette hele arbeidsdag. Da må man begynne å bli bekymret for sin kropp. Det er best å gå på tur eller sykle. Noen trener på sportklubb eller svømmer i UNK. Godt UNK er også en av de viktigste momentene. Vi forlenger livet når vi smiler eller tenker positivt. Det virker veldig UNK å ha en god helse og leve lenge. Livet er spent. Man kommer i verden.

Essay by Russian native speaker

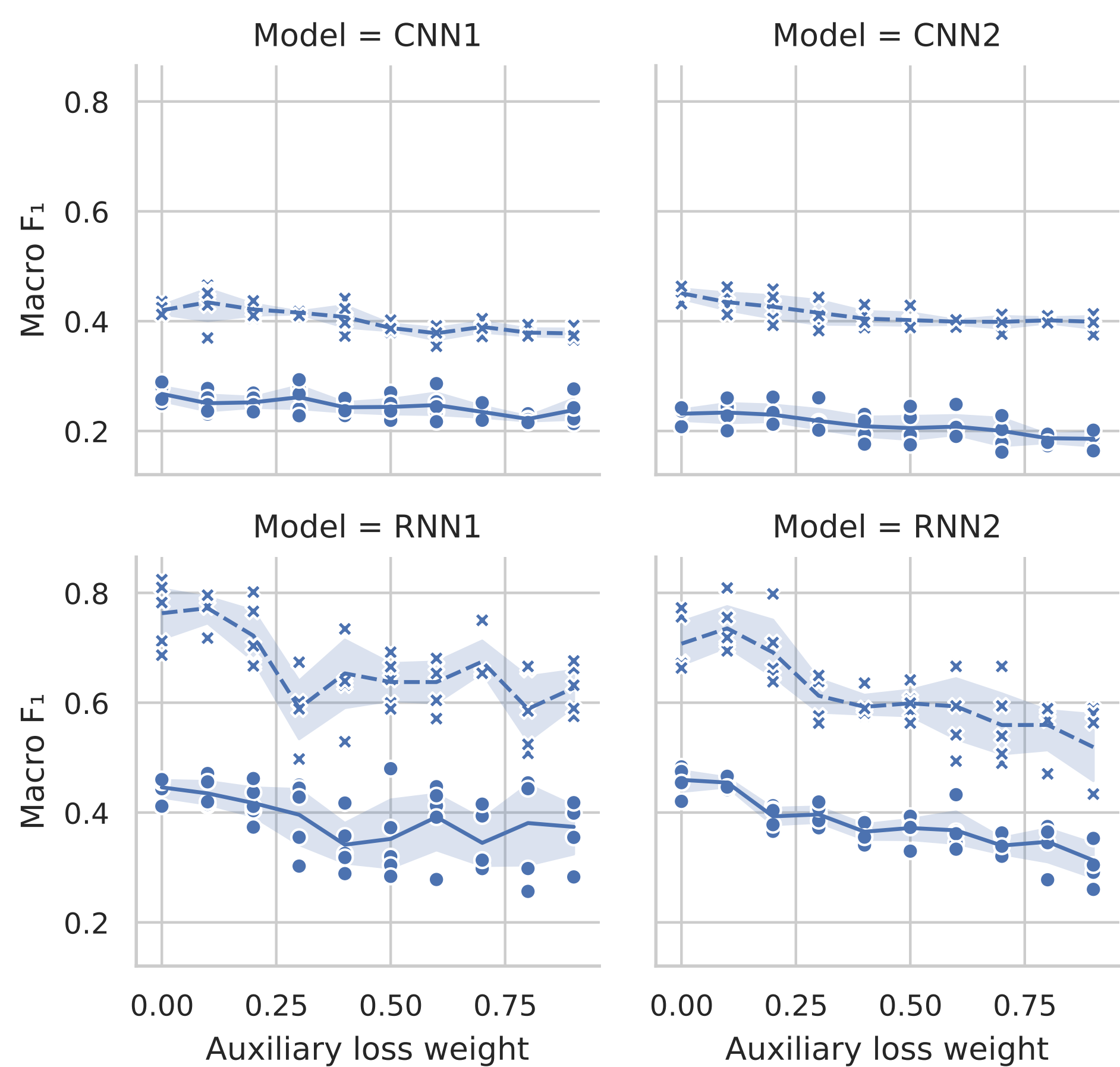
NLI

Model	Macro F ₁	Micro F ₁
Mean	0.520	0.537
Max	0.401	0.390
Attn	0.447	0.480
+POS Mean	0.467	0.480
+POS Max	0.406	0.431
+POS Attn	0.454	0.463

Multi-task Learning: Best settings

Hyperparameter	RNN1	RNN2
Word embeddings	Dynamic	
Embedding size	100	
RNN cell	GRU	
Pooling method	Attention	
Bidirectional	Yes	
Embedding init	Random	Pre-trained
Input representation	Tokens+UPOS	Tokens

Multi-task Learning: Loss curves



Held-out set results

Model	All labels		Collapsed labels	
	Macro	Micro	Macro	Micro
Majority	0.045	0.187	0.127	0.341
SVR BOW	0.231	0.285	0.420	0.602
SVR POS	0.271	0.350	0.422	0.602
RNN1	0.291	0.439	0.478	0.724
RNN2	0.388	0.480	0.511	0.724
Multi-RNN1	0.266	0.398	0.509	0.707
Multi-RNN2	0.356	0.447	0.443	0.724

Error Analysis

A2	1	2	0	0	0	0	0	eng	4	1	0	0	3	0	0
A2/B1	1	9	9	0	0	0	0	pol	6	4	1	0	11	0	0
B1	0	6	10	5	2	0	0	rus	0	0	4	0	9	0	3
B1/B2	0	2	3	7	8	2	0	som	0	1	0	4	6	0	2
B2	0	0	2	9	25	4	0	spa	4	1	1	1	10	0	0
B2/C1	0	0	0	1	5	3	0	deu	11	2	0	0	3	6	0
C1	0	0	0	0	3	4	0	vie	1	1	2	3	8	0	10
A2A2/B1B1B1/B2B2B2/C1C1								eng pol rus som spa deu vie							

Conclusion

- AES task is best modeled as regression for ASK corpus.
- mean-over-time BiGRU model performed the best at NLI task.
- Auxiliary loss weight of 0.1 is best suited for joint modeling of AES and NLI tasks.
- Pretrained embeddings achieve the best results.