International Conference on Artificial Intelligence in Education Technology (AIET 2021)

TIM SCHLIPPE & JÖRG SAWATZKI

CROSS-LINGUAL AUTOMATIC SHORT ANSWER GRADING

Wuhan, China
July 4, 2021

SILICON SURFER
iu INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

AGENDA

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

MOTIVATION

MOTIVATION

SUSTAINABLE DEVELOPMENT GOALS

Sources: United Nations: Sustainable Development Goals: 17 Goals to Transform our World (2021); OpenClipart-Vectors/154119/Pixabay.

MOTIVATION

SUSTAINABLE DEVELOPMENT GOALS

Sources: United Nations: Sustainable Development Goals: 17 Goals to Transform our World (2021); OpenClipart-Vectors/154119/Pixabay; Statista: The Most Spoken Languages Worldwide in 2019 (2020).

MOTIVATION

CROSS-LINGUAL
AUTOMATIC SHORT ANSWER GRADING

Graphic Source: Schlippe & Sawatzki (2021a).

MOTIVATION

Question	What is a variable?
Model answer	A location in memory that can store a value.
Answer 1	Eine Stelle im Speicher, die einen Wert speichern kann.
Grading: Answer 1	**5 of 5 points**
Answer 2	变量可以是整数，也可以是程序中的字符串。
Grading: Answer 2	**2 of 5 points**

Graphic Source: Custom Depiction.

02

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

CROSS-LINGUAL AUTOMATIC SHORT ANSWER GRADING

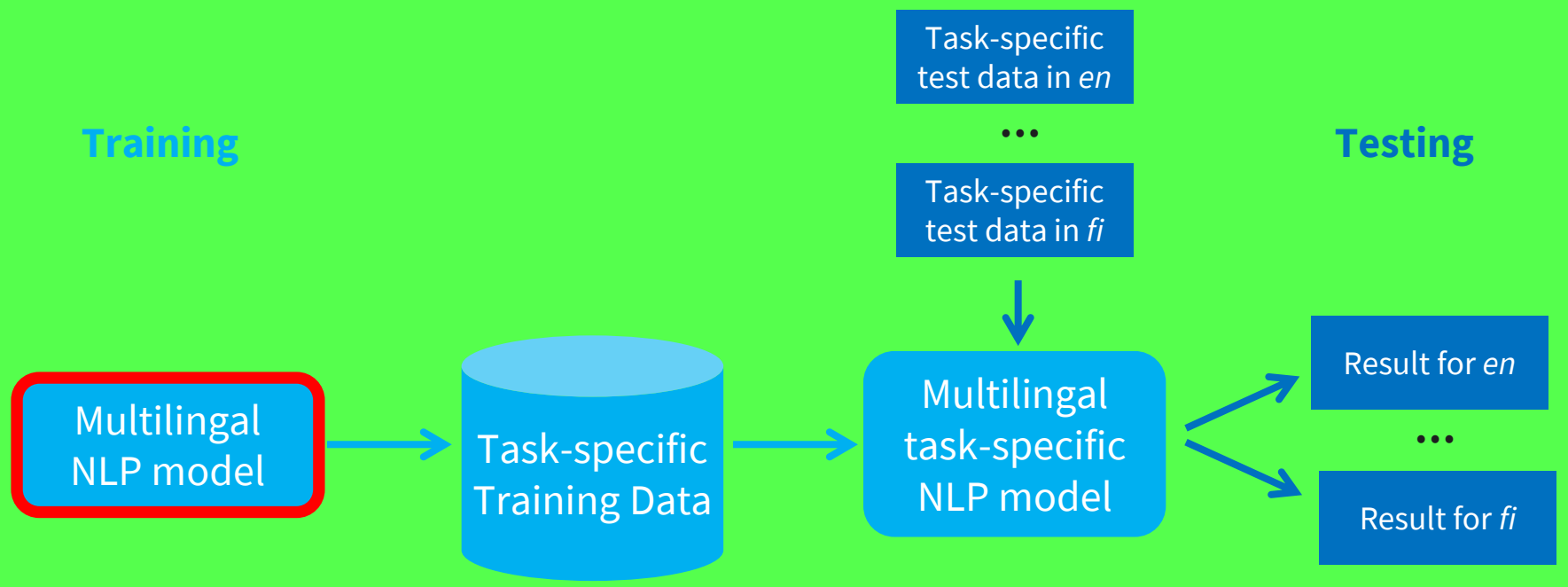
CROSS-LINGUAL AUTOMATIC SHORT ANSWER GRADING

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

MULTILINGUAL NLP MODELS

Transfer learning
Cross-lingual transfer

e.g., Multilingual BERT (Devlin et al., 2019; Pires et al., 2019), RoBERTa (Liu et al., 2019), XLM-R (Conneau, 2018)



Graphic Source: Custom Depiction.

CROSS-LINGUAL AUTOMATIC SHORT ANSWER GRADING

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

MULTILINGUAL NLP MODELS

Transfer learning

Cross-lingual transfer

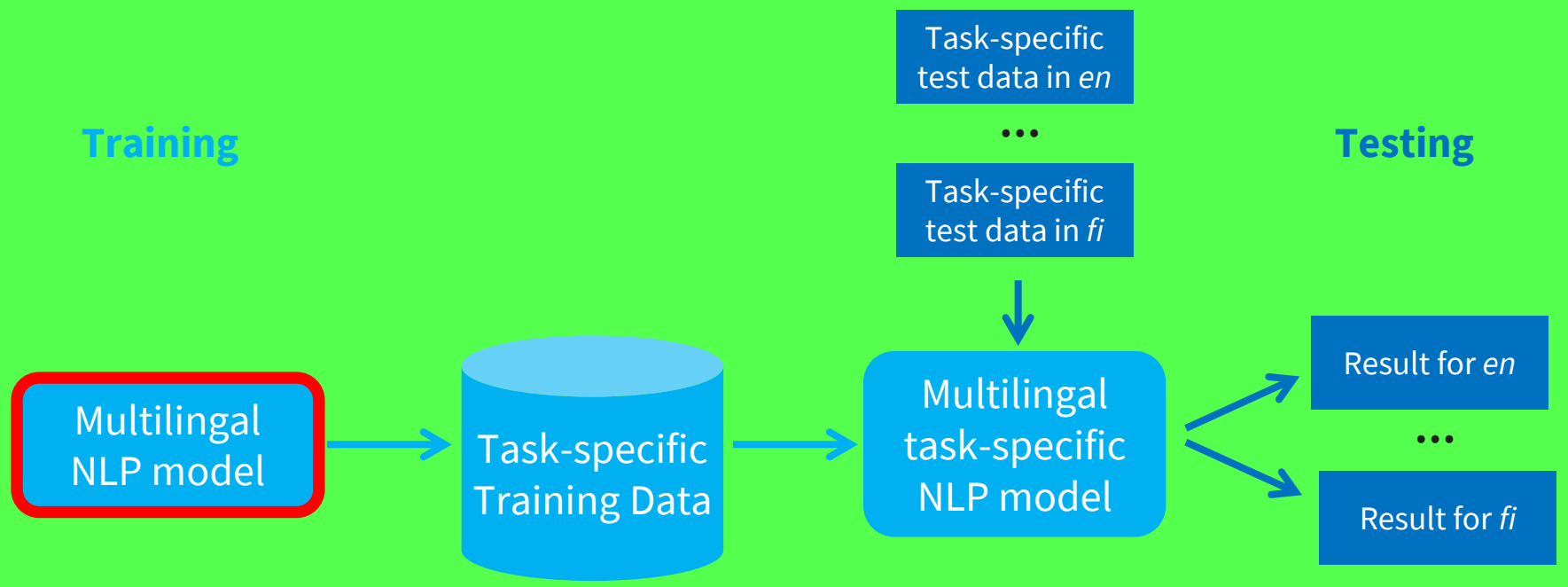
e.g., Multilingual BERT

Multilingual BERT

Pires et al., 2019),

RoBERTa (Liu et al., 2019),

XLM-R (Conneau, 2018)



Graphic Source: Custom Depiction.

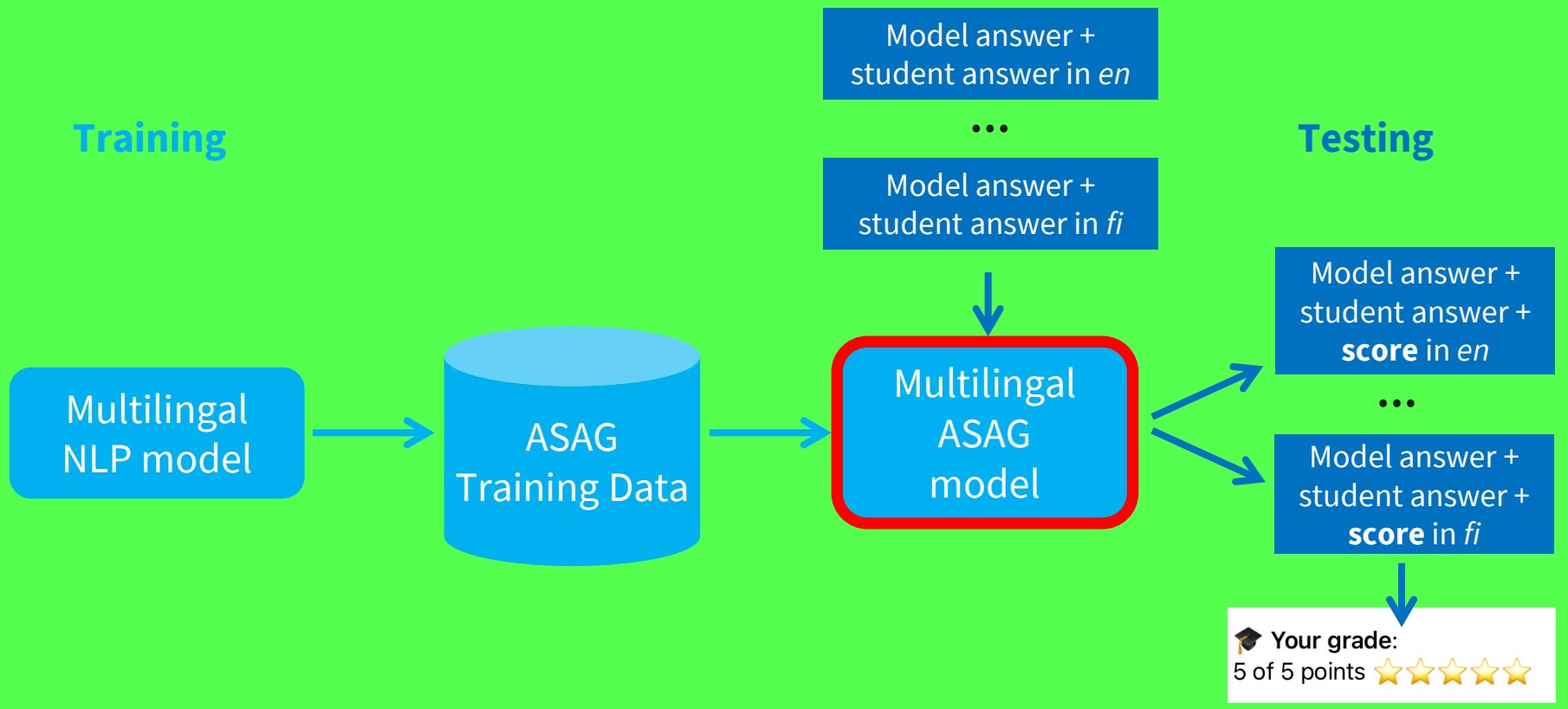
CROSS-LINGUAL AUTOMATIC SHORT ANSWER GRADING

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

AUTOMATIC SHORT ANSWER GRADING

Deep learning

e.g., (Burrows et al., 2014; Camus & Filighera, 2020; Sawatzki et al., 2021; Schlippe & Sawatzki, 2021b)



Graphic Source: Custom Depiction.

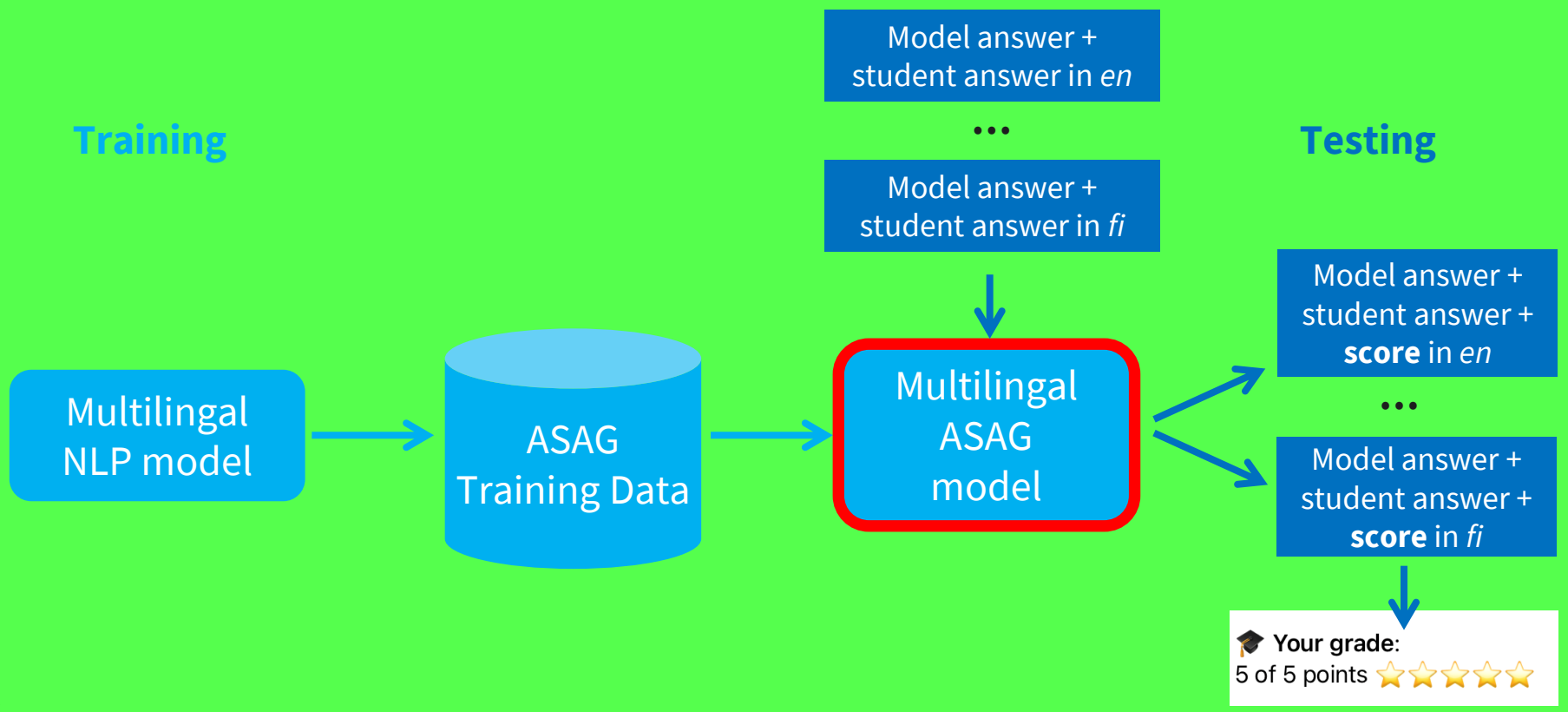
CROSS-LINGUAL AUTOMATIC SHORT ANSWER GRADING

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

AUTOMATIC
SHORT ANSWER
GRADING

Deep learning

Cross-lingual Automatic Short Answer Grading



Graphic Source: Custom Depiction.

03

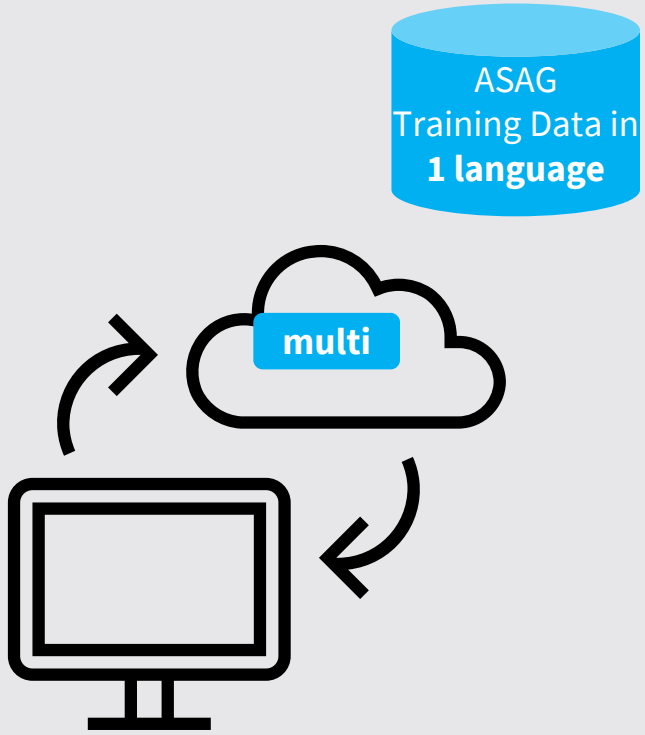
INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

EXPERIMENTAL SETUP

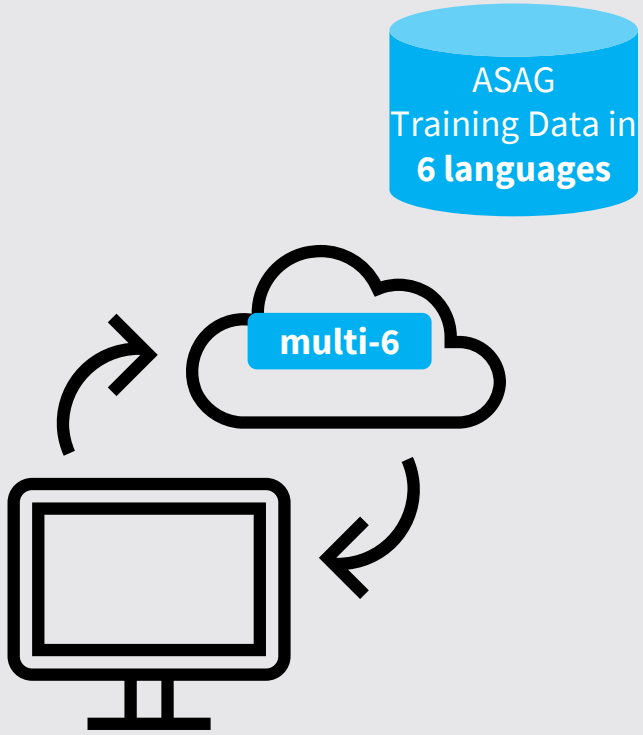
EXPERIMENTAL SETUP

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

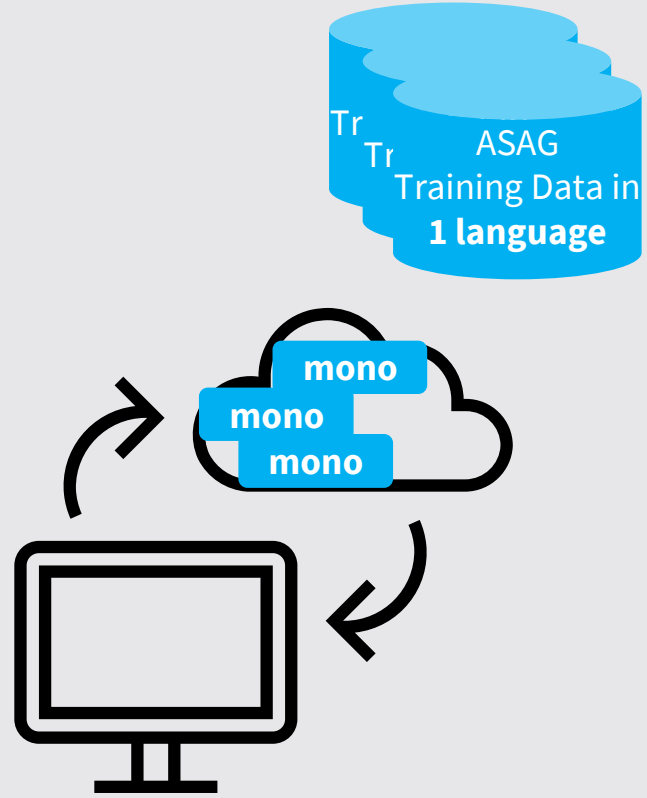
MULTILINGUAL MODEL
fine-tuned on 1 single language



MULTILINGUAL MODEL
fine-tuned on 6 languages



MONOLINGUAL MODELS



Graphic Source: Custom Depiction.

EXPERIMENTAL SETUP

Question	What is a variable?
Model answer	A location in memory that can store a value.
Example: Answer 1	A variable is a location in memory where a value can be stored.
Grading: Answer 1	5 of 5 points
Example: Answer 2	Variable can be an integer or a string in a program.
Grading: Answer 2	2 of 5 points

Translation Model

Google's Neural Machine Translation System
(Wu et al., 2016; Aiken, 2019)

Translation Model

Question	Was ist eine Variable?
Model answer	Eine Stelle im Speicher, die einen Wert speichern kann.
Example: Answer 1	Eine Variable ist ein Ort im Speicher, an dem ein Wert gespeichert werden kann.
Grading: Answer 1	5 of 5 points
Example: Answer 2	Eine Variable kann in einem Programm ein Integer oder ein String sein.
Grading: Answer 2	2 of 5 points

Question	什么是变量？
Model answer	内存中可以存储值的位置。
Example: Answer 1	变量是内存中可以存储值的位置。
Grading: Answer 1	5 of 5 points
Example: Answer 2	变量可以是整数，也可以是程序中的字符串。
Grading: Answer 2	2 of 5 points

Sources: Schlippe & Sawatzki (2021b); Mohler et al. (2011).

04

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

EXPERIMENTS AND RESULTS

EXPERIMENTS AND RESULTS

	multi+ en	multi+ de	multi+ nl	multi+ jp	multi+ zh	multi+ fi	multi+ 6	mono
en	0.45	0.61	0.64	0.68	0.63	0.63	**0.43**	0.43
ceb	0.70	0.73	0.72	0.68	0.72	0.71	**0.63**	-
sv	0.63	0.67	0.68	0.73	0.72	0.68	**0.48**	-
de	0.64	0.51	0.67	0.70	0.70	0.65	0.46	**0.45**
fr	0.61	0.66	0.64	0.67	0.70	0.67	**0.54**	-
nl	0.62	0.64	0.52	0.70	0.73	0.67	**0.45**	0.47
ru	0.68	0.73	0.83	0.74	0.75	0.78	**0.52**	-
it	0.62	0.65	0.72	0.71	0.73	0.70	**0.52**	-
es	0.61	0.68	0.76	0.68	0.72	0.65	**0.49**	-
pl	0.62	0.71	0.77	0.69	0.72	0.68	**0.51**	-
vi	0.71	0.72	0.84	0.77	0.73	0.71	**0.52**	-
jp	0.66	0.70	0.73	0.49	0.63	0.71	**0.44**	0.53
zh	0.63	0.71	0.77	0.69	0.50	0.79	**0.41**	0.44
ar	0.72	0.78	0.85	0.78	0.76	0.76	**0.59**	-
uk	0.65	0.70	0.82	0.73	0.73	0.75	**0.54**	-
pt	0.59	0.67	0.75	0.69	0.73	0.69	**0.50**	-
fa	0.64	0.66	0.71	0.67	0.70	0.69	**0.56**	-
ca	0.64	0.70	0.74	0.70	0.76	0.67	**0.53**	-
sr	0.69	0.81	0.83	0.76	0.79	0.86	**0.56**	-
id	0.66	0.68	0.69	0.70	0.79	0.63	**0.49**	-
no	0.63	0.69	0.65	0.75	0.71	0.69	**0.45**	-
ko	0.70	0.70	0.76	0.66	0.66	0.67	**0.58**	-
fi	0.69	0.79	0.77	0.77	0.73	0.52	0.47	**0.45**
hu	0.69	0.76	0.81	0.72	0.76	0.69	**0.54**	-
cs	0.62	0.77	0.82	0.72	0.78	0.71	**0.51**	-
sh	0.66	0.77	0.79	0.74	0.78	0.79	**0.53**	-

Mean Absolute Error out of 5 points

Graphic Sources: Schlippe & Sawatzki (2021b).

EXPERIMENTS AND RESULTS

0.75 POINTS
HUMAN GRADER VARIABILITY

	multi+ en	multi+ de	multi+ nl	multi+ jp	multi+ zh	multi+ fi	multi+ 6	mono
en	0.45	0.61	0.64	0.68	0.63	0.63	**0.43**	0.43
ceb	0.70	0.73	0.72	0.68	0.72	0.71	**0.63**	-
sv	0.63	0.67	0.68	0.73	0.72	0.68	**0.48**	-
de	0.64	0.51	0.67	0.70	0.70	0.65	0.46	**0.45**
fr	0.61	0.66	0.64	0.67	0.70	0.67	**0.54**	-
nl	0.62	0.64	0.52	0.70	0.73	0.67	**0.45**	0.47
ru	0.68	0.73	0.83	0.74	0.75	0.78	**0.52**	-
it	0.62	0.65	0.72	0.71	0.73	0.70	**0.52**	-
es	0.61	0.68	0.76	0.68	0.72	0.65	**0.49**	-
pl	0.62	0.71	0.77	0.69	0.72	0.68	**0.51**	-
vi	0.71	0.72	0.84	0.77	0.73	0.71	**0.52**	-
jp	0.66	0.70	0.73	0.49	0.63	0.71	**0.44**	0.53
zh	0.63	0.71	0.77	0.69	0.50	0.79	**0.41**	0.44
ar	0.72	0.78	0.85	0.78	0.76	0.76	**0.59**	-
uk	0.65	0.70	0.82	0.73	0.73	0.75	**0.54**	-
pt	0.59	0.67	0.75	0.69	0.73	0.69	**0.50**	-
fa	0.64	0.66	0.71	0.67	0.70	0.69	**0.56**	-
ca	0.64	0.70	0.74	0.70	0.76	0.67	**0.53**	-
sr	0.69	0.81	0.83	0.76	0.79	0.86	**0.56**	-
id	0.66	0.68	0.69	0.70	0.79	0.63	**0.49**	-
no	0.63	0.69	0.65	0.75	0.71	0.69	**0.45**	-
ko	0.70	0.70	0.76	0.66	0.66	0.67	**0.58**	-
fi	0.69	0.79	0.77	0.77	0.73	0.52	0.47	**0.45**
hu	0.69	0.76	0.81	0.72	0.76	0.69	**0.54**	-
cs	0.62	0.77	0.82	0.72	0.78	0.71	**0.51**	-
sh	0.66	0.77	0.79	0.74	0.78	0.79	**0.53**	-

Mean Absolute Error out of 5 points

Graphic Sources: Schlippe & Sawatzki (2021b).

EXPERIMENTS AND RESULTS

	multi+ en	multi+ 6	rel. improvement
en	0.45	0.43	4.4%
de	0.64	0.46	28.1%
nl	0.62	0.45	27.4%
jp	0.66	0.44	33.3%
zh	0.63	0.41	34.9%
fi	0.69	0.47	31.9%

Mean Absolute Error
out of 5 points
(Human grader variability: 0.75)

➔ Up to 35% improvement by adding more languages

Graphic Sources: Schlippe & Sawatzki (2021b).

EXPERIMENTS AND RESULTS

	multi+ L_{target}	multi+ 6	rel. improvement
en	0.45	0.43	4.4%
de	0.51	0.46	9.8%
nl	0.52	0.45	13.5%
jp	0.49	0.44	10.2%
zh	0.50	0.41	18.0%
fi	0.52	0.47	9.6%

Mean Absolute Error
out of 5 points
(Human grader variability: 0.75)

➔ Even with data in the target language, adding the 5 languages provides improvements of up to 18%

Graphic Sources: Schlippe & Sawatzki (2021b).

05

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

CONCLUSION AND FUTURE WORK

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Conclusion

- Potential of cross-lingual automatic short answer grading
- Analysis on 26 languages
- Mean Absolute Errors (MER) between 0.41 and 0.72 points out of 5 points
- Less discrepancy than 2 human graders
- Augmenting training data with machine translated task-specific data for fine-tuning improves multilingual models
- Results experimentally

CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

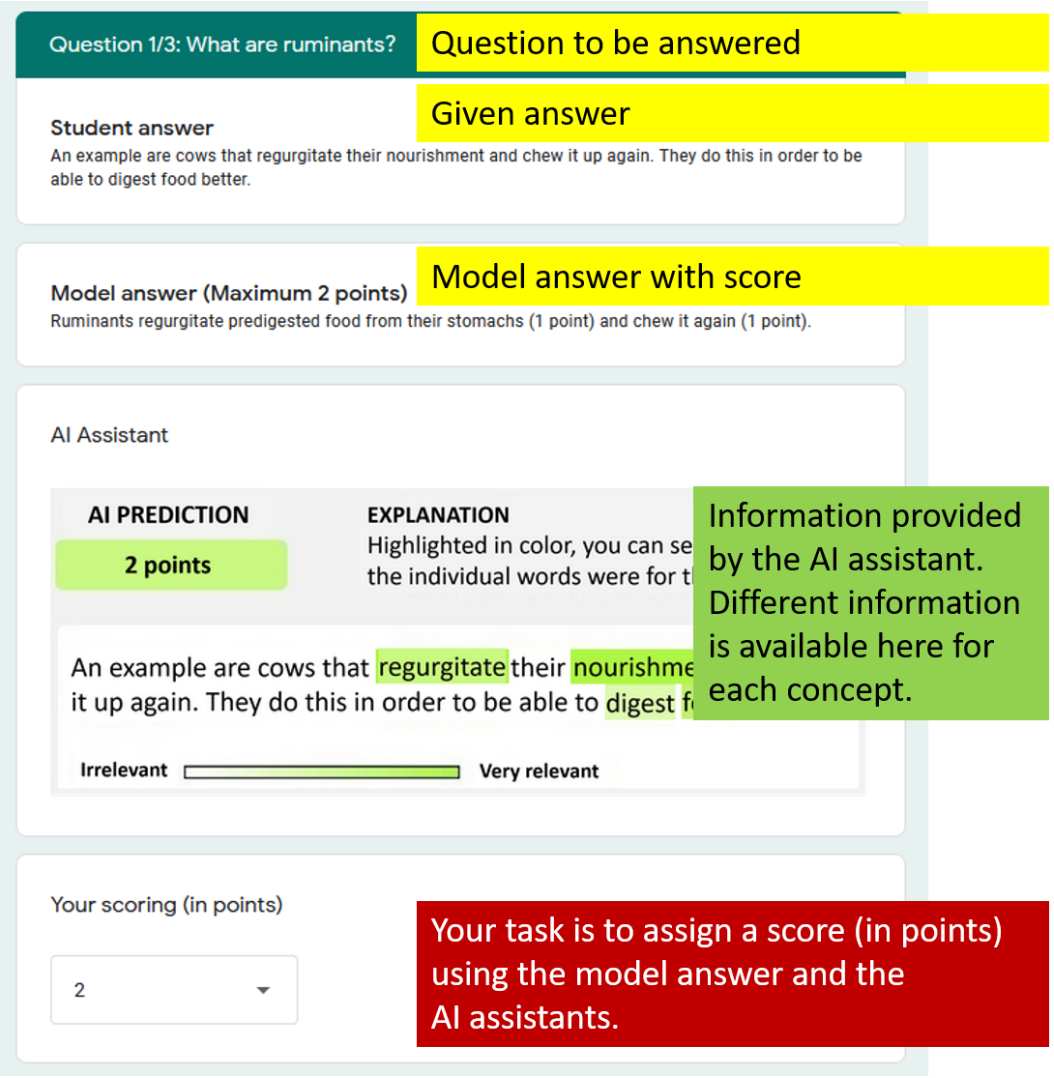
Conclusion

- Potential of cross-lingual automatic short answer grading
- Analysis on 26 languages
- Mean Absolute Errors (MER) between 0.41 and 0.72 points out of 5 points
- Less discrepancy than 2 human graders
- Augmenting training data with machine translated task-specific data for fine-tuning improves multilingual models
- Results experimentally

Future Work

- Extension to other languages
- Integration and application for online exams
- Interactive training programs for exam preparation (Schlippe & Sawatzki, 2021a)
- Address the issue of explainability to provide better support to human graders

FUTURE WORK



https://tinyurl.com/AI-grade

Try it out and support us.

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

THANK YOU

Tim Schlippe
tim.schlippe@iu.org

REFERENCES

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Literature

- **United Nations: Sustainable Development Goals: 17 Goals to Transform our World (2021)**: https://www.un.org/sustainabledevelopment/sustainable-development-goals
- **Statista: The Most Spoken Languages Worldwide in 2019 (2020)**: https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide
- **Schlippe, T. & Sawatzki, J. (2021a)**: *AI-based Multilingual Interactive Exam Preparation*. In: Proceedings of The Learning Ideas Conference 2021 (14th annual conference), ALICE - Special Conference Track on Adaptive Learning via Interactive, Collaborative and Emotional Approaches, New York, New York, 14-18 June 2021. In: Advances in Intelligent Systems and Computing, Springer, 2021.
- **Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019)**: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In: NAACL-HLT, Minneapolis, Minnesota.
- **Pires, T., Schlinger, E., & Garrette, D. (2019)**: *How Multilingual is Multilingual BERT?* In: ACL. Florence, Italy.
- **Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019)**: *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv:1907.11692.

Images

- **Images provided by OpenClipart-Vectors/154119/Pixabay**. (https://pixabay.com/vectors/international-project-world-154119 [last access: 16.16.2021])
- **Coursebook (NLP and Computer Vision)**: IU International University of Applied Sciences. Version No.: 001-2020-1211.

REFERENCES

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Literature

- **Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2018)**: *Unsupervised Cross-lingual Representation Learning at Scale*. arXiv:1911.02116.
- **Burrows, S., Gurevych, I., & Stein, B. (2014)**: *The Eras and Trends of Automatic Short Answer Grading*. In: IJAIED 25, 60–117.
- **Camus, L. & Filighera, A. (2020)**: *Investigating Transformers for Automatic Short Answer Grading*. AIED. Cyberspace.
- **Sawatzki, J., Schlippe, T., & Benner-Wickner, M. (2021)**: *Deep Learning Techniques for Automatic Short Answer Grading: Predicting Scores for English and German Answers*. In: AIET, Wuhan, China.
- **Schlippe, T. & Sawatzki, J. (2021b)**: *Cross-Lingual Automatic Short Answer Grading*. In: AIET, Wuhan, China.
- **Mohler, M., Bunescu, R., & Mihalcea, R. (2011)**: Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 752–762. Association for Computational Linguistics, Portland, Oregon, USA.