Adversarial Paper Pläne

ASAP-Datensätze

- Englisch (original, orig)
- Englisch (kleinerer Datensatz, orig300)
- Englisch (Crowdworkers, en)
- Deutsch (de)
- Französisch (fr)
- Spanisch (es)
- Chinesisch (zh)

Horbach, A., Pehlke, J., Laarmann-Quante, R., & Ding, Y. (2023). Crosslingual Content Scoring in Five Languages Using Machine-Translation and Multilingual Transformer Models. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-023-00370-1

Methoden

Methode	Paper	Wer übernimmt?	Was soll gemacht werden?
Character N-Grams	Ding et al. (2020)	Noemi	• Basierend auf den echten Antworten • Originalcode von https://github.com/catalpa-cl/adversarials ausprobieren • Mit eigenem Code aus dem Kurs vergleichen: Unterschiede? • Auf die anderen Sprachen anwenden
Word N-Grams	Ding et al. (2020)	Noemi?	
Shuffle	Ding et al. (2020)		
Content Burst (Prompt und/oder Antworten)	Ding et al. (2020)	Christopher	• Nomen aus dem Prompt shufflen: Originalcode von https://github.com/catalpa-cl/adversarials für „Content Burst“ ausprobieren und für alle Sprachen anpassen (anderer POS-Tagger!)
Prompt N-Grams	eigene Erweiterung	Luca	• Beliebige N-Gramme aus dem Prompt shufflen
Adjektive + Adverbien einfügen	Filighera et al. (2023)	Adjektive: Vitalia Adverbien: Alona	• Adverbien + Adjektive in falsche Antworten einfügen • Originalcode verwendbar? https://github.com/SebOchs/adversarial_insertions • **Adjektive**: Wie umgehen mit korrekter Flexion bei anderen Sprachen? • **Adverbien**: Wo stehen in den anderen Sprachen natürlicherweise die Adverbien?

Benötigte Ressourcen

	Englisch	Deutsch	Französisch	Spanisch
Character N-Grams	• Generisches Korpus: Brown • alle echten Antworten	• Generisches Korpus • alle echten Antworten	• Generisches Korpus • alle echten Antworten	• Generisches Korpus • alle echten Antworten
Word N-Grams	• Generisches Korpus • alle echten Antworten	• Generisches Korpus • alle echten Antworten	• Generisches Korpus • alle echten Antworten	• Generisches Korpus • alle echten Antworten
Shuffle	• echte richtige Antworten	• echte richtige Antworten	• echte richtige Antworten	• echte richtige Antworten
Prompt/Content Burst	• POS-Tagger • Prompt	• POS-Tagger • Prompt	• POS-Tagger • Prompt	• POS-Tagger • Prompt
Adjektive + Adverbien	• POS-Tagger • Generisches Korpus • echte falsche Antworten	• POS-Tagger • Generisches Korpus • echte falsche Antworten	• POS-Tagger • Generisches Korpus • echte falsche Antworten	• POS-Tagger • Generisches Korpus • echte falsche Antworten

Synergieeffekte nutzen: Es sollten innerhalb einer Sprache die gleichen Ressourcen (generisches Korpus, POS-Tagger) genutzt werden! spaCy sollte fürs POS-Tagging aller Sprachen funktionieren

Wie soll das Ergebnis aussehen?

Jeweils pro Prompt, pro Datensatz:

1. txt-Datei mit generierten Adversarial-Antworten; 1 Antwort pro Zeile
   - Dateibenennung: Adversarial_*Datensatz*_Prompt*X*_Methode.txt, z.B. Adversarial_de_prompt1_word_1grams.txt
2. Adversarial Rejection Rate (einheitliches Skript zur Berechnung liefert RLQ)
   - Finale Werte hier eintragen: https://cryptpad.fr/sheet/#/2/sheet/edit/Pg3cFGms7a8pbw9nTvT3ZKra/

	Englisch (orig)			**Englisch klein (orig300)**			**Englisch (en)**			**Deutsch (de)**			**Französisch (fr)**			**Spanisch (es)**		
	Prompt 1	Prompt 2	Prompt 10	Prompt 1	Prompt 2	Prompt 10	Prompt 1	Prompt 2	Prompt 10	Prompt 1	Prompt 2	Prompt 10	Prompt 1	Prompt 2	Prompt 10	Prompt 1	Prompt 2	Prompt 10
ARR																		

Dokumentation

- Es muss keine ganze Hausarbeit geschrieben werden, sondern nur etwas zu **Methode**, **Ergebnis** und **Diskussion** bezogen auf die eigene Methode
  - inkl. geeignete Visualisierung der Ergebnisse
- Zusätzlich: nach **Related Work** Ausschau halten und ggf. im Methoden-Teil mit einbauen
- Wichtig: auf Englisch schreiben
- Gemeinsames LaTeX-Dokument in Overleaf: https://www.overleaf.com/1292182874wgvfzgdysmjz#11ac62
  - Wer sich zu unsicher fühlt in LaTeX kann auch zunächst in Word schreiben
- Die Beschreibungen sollten für die PL zunächst ausführlicher ausfallen, für das finale Paper wird dann gekürzt