Argumentzulänglichkeitsbewertung mittels Konklusionsregenerierung und semantischer Ähnlichkeit

Projektbericht Argument Mining

Humanwissenschaftliche Fakultät
Department Linguistik

vorgelegt von: Simon Bross

Matrikelnummer: 809648

Berlin, den 24. Oktober 2024

1. Einleitung

Gurcke et al. (2021) untersuchten in ihrem Papier „Assessing the Sufficiency of Arguments through Conclusion Generation“, inwieweit die computergestützte Bewertung von Argumentzulänglichkeit von der automatischen Regenerierung der Konklusion anhand der Prämissen profitiert. Die wenigen vorangegangenen Arbeiten, die sich mit der Qualitätsdimension der *argument sufficiency* beschäftigten, ignorierten jedoch jene Beziehung zwischen Prämissen und Konklusionen und sahen die Aufgabe als ein Textklassifikationsproblem an, welches sie mit Hilfe von Convolutional Neural Networks lösten. Mit ihrer Arbeit versuchten Gurcke et al. nicht nur diese Ansätze zu erweitern und dabei der Beziehung zwischen Prämissen und Konklusion gerecht zu werden, sondern auch einen Beitrag zur Verbesserung des aktuellen state-of-the-art (Macro F1 Wert von 0.827; menschliche Performanz: Macro F1 Wert von 0.887) zu leisten (vgl. ebd., S. 67-69).

In Anlehnung an die Argumentqualitätsforschung, in der die Dimension der *(local) sufficiency* solche Argumente als zulänglich betrachtet, deren Prämissen in ihrer Gesamtheit genügend Unterstützung für ein rationales Ableiten der Konklusion liefern (vgl. Wachsmuth et al., 2017, S. 182), formulierten die Autoren ihre Arbeitshypothese, dass einzig zulängliche Argumente eine Ableitung der Konklusion aus den Prämissen ermöglichen (vgl. Gurcke et al., 2021, S. 68). Zur Untersuchung stützten sie sich dabei auf den Vergleich zwischen der ursprünglich aufgestellten und einer von einem Sprachmodell (BART) (re-)generierten Konklusion, die einem weiteren Sprachmodell (RoBERTa) mit Klassifikationslayer bei der Klassifizierung von Argumentzulänglichkeit respektive Argumentunzulänglichkeit unterstützen soll. Die Hypothese wirft damit die Frage auf, ob das im Sprachmodell kodierte Wissen und seine Inferenzfähigkeiten für diese Aufgabe wirksam angewendet werden können.

Obwohl Gurcke et al. mit ihren Ergebnissen den state-of-the-art verbessern, sich weiter an die menschliche Performanz annähern und teilweise sogar überteffen konnten (Macro F1 Wert Gurcke et al.: 0.885, menschliche Performanz: 0.876), war in ihrem Verfahren nicht transparent, nach welchen Kriterien beziehungsweise Merkmalen ein Argument als un-/zulänglich klassifiziert wird. Dies ist auf den Black-Box-Charakter neuronaler Netze zurückzuführen, der die Nachvollziehbarkeit einer internen Entscheidung verhindert.

An dieser Stelle setzt das vorliegende Projekt an, in dem versucht wird, die Arbeit von Gurcke et al. teilweise zu replizieren, jedoch mit dem Ziel, den Klassifizierungsprozess transparenter zu gestalten. Die Arbeitshypothese und der erste Verfahrensschritt von Gurcke et al. - die Konklusionsregenerierung mittels BART - werden dazu übernommen, anstelle der Klassifikation durch ein neuronales Netz experimentiert dieses Projekt hingegen mit einem semantischen Vergleichswert zwischen der ursprünglichen und generierten Konklusion, der als feature für einen kleinen in Python selbstgeschriebenen Klassifikator fungiert. Auf dieser Basis wird ein Argument nun als zulänglich klassifziert, wenn seine Konklusion in einem semantisch ähnlichen Maß - sprich oberhalb eines intern aus den Daten abgeleiteten Schwellenwerts - im Vergleich zur Originalkonklusion durch BART alleinig aus den Prämissen regeneriert werden konnte.

2. Daten

Als Datengrundlage für das Projekt diente, wie auch bei Gurcke et al., die zweite Version des Korpus *Argument Annotated Essays* von Stab und Gurevych (2017). Er enthält 402 komplette studentische Aufsätze mit Annotationen der Argumentationsstruktur. Von den insgesamt 1029 singulären Argumenten, die in der Originalversion enthalten waren, sind in der XML-geparsten Version des Korpus, die für dieses Projekt aus dem GitHub-Repository von Gurcke (2021) extrahiert wurde, noch 982 vorhanden. Der Datensatz ist unausgeglichen und besteht insgesamt aus ca. 66% als zulänglich (Label: 1.0) und 34% als unzulänglich (Label: 0.0) annotierten Argumenten.

Die Argumente des Datensatzes wurden im Preprocessing um eine Variante erweitert, bei der die Konklusion mit einem '<mask>' Token maskiert wurde. Die maskierten Argumente dienen als Inputs für das BART-Modell, das durch Abgleich mit den originalen Konklusionen als Labels lernt, eine in den Kontext passende Konklusionen alleinig aus den Prämissen zu generieren. Für die Konklusionsregenerierung und Argumentzulänglichkeitsklassifizierung wurden 80% der Daten für das Training und 20% für das Testing verwendet, sowie 5-fold cross validation durchgeführt. Während des Finetunings von BART wurden von den Trainingsdaten 15% intern zur Validierung des Modells verwendet.

3. Zweischrittiges Verfahren

3.1 Konklusionsregenerierung

3.1.1 Finetuning

Für die Aufgabe der Konklusionsregenerierung wurde zunächst das Sprachmodell BART in der large-Variante (400 Mrd. Parameter) auf die in 2. genannten Daten finegetuned. Dabei wurden seine denoising-Fähigkeiten und bidirektionale Aufmerksamkeit ausgenutzt, um das <mask> Token im maskierten Argument auf eine Konklusion abzubilden, die in den jeweiligen Kontext passt - das Modell lernte durch das Finetuning folglich, das Token durch Text zu ersetzen (*text infilling*).

Probleme zeigten sich hauptsächlich bei der Festlegung der Hyperparameter, die starken Einfluss auf die Generierungsqualität des Modells nehmen. Repetitive und unsinnige Generierungen wie „Additional Deluxe Deluxe Deluxe Revised Deluxe Deluxe Voting Deluxe Deluxe Spani...“ gegenüber der Originalkonklusion „through cooperation, children can learn about interpersonal skills which are significant in the future life of all students“ waren Indiz für ein anfängliches Overfitting des Modells, das durch die richtigen Hyperparametereinstellungen gelöst wurde. Bei der Festlegung orientierte ich mich an den Einstellungen von Gurcke et al. - falls im Papier vermittelt - und experimentierte zusätzlich durch manuellen Abgleich der Qualität der regenerierten Konklusionen.

Folgende Hyperparameter wurden verwendet:

1. Learning Rate: 5×10^{-6} mit Learning Rate Scheduler (Cosine Decay) zur Trainigsstabilisierung
2. Optimizer: Adam Optimizer
3. Loss function: Sparse Categorical Cross Entropy
4. Epochs: 10, mit early stopping callback (patience = 2) und Speicherung der Gewichte der besten Epoche (anhand Validierungsloss)
5. Batch Size: 5

Generierungen mit diesen Einstellungen wiesen trotz vereinzelter grammatikalischer Fehler und redundanter Zeichen oder Wörter eine deutlich bessere Qualität (Kohärenz und Kohäsion) auf und waren teilweise sogar gehaltvoller als die Originalkonklusion: „One of the advantages of studying in a foreign country is that they will be exposed to different culture [sic]. For example, they get exposure to a different educational system. They will meet new professors and new classmates which makes the academic experience different from that in their home country. “ (generiert) gegenüber „the new academic experience that the students can obtain at the institution where they are pursuing their studies “ (Original).

3.1.2 Inferenz

Wie in den Beispielgenerierungen erkennlich ist, sind die generierten Konklusionen im Vergleich zu den originalen oftmals länger beziehungsweise bestanden aus mehreren Sätzen, wohingegen die Konklusionen aus den Studentenessays immer nur aus einem Satz bestanden. Um die semantische Ähnlichkeit nicht zu sehr durch äußerst ungleiche Konklusionen zu verfälschen, wurde deshalb aus den Generierungen nur der erste Satz extrahiert. Ein generelles Forcieren zur Generierung von Konklusionen, die nur aus einem Satz bestehen, ist offenbar nicht möglich.

Folgende Parameter wurden zur Regenerierung der Konklusionen verwendet (`model.generate()`):

1. max_length: 70 (Tokens; orientiert an der maximalen Konklusionslänge im Korpus)
2. min_length: 20: (Tokens; orientiert an der minimalen Konklusionslänge im Korpus)
3. no_repeat_ngram_size: 2 (Das Wiederholen von Bigrammen wurde unterbunden)
4. num_beams: 5 Beam Searches wurden durchgeführt
5. temperature: 0,7 (um die Generierungen konservativer zu machen)
6. repetition_penalty: 1,8 (verhindert redundante, aufeinanderfolgende Wortwiederholungen)

3.2 Klassifizierung von Argumentzulänglichkeit

Mit Paaren bestehend aus generierter und originaler Konklusion wurde ein kleiner Klassifikator trainiert, der die beiden Konklusionen in sentence embeddings transferiert und zwischen ihnen die Kosinusähnlichkeit bestimmt. Die Werte der Kosinusähnlichkeit wurden von ihrem ursprünglichen Wertebereich $[-1, 1]$ auf einen Bereich von $[0, 1]$ reskaliert - unter der Annahme, dass sie dadurch die Wahrscheinlichkeit repräsentieren, dass ein Argument zur Klasse 1 (zulänglich) gehört. Zur Bestimmung des optimalen Schwellenwertes des Wahrscheinlichkeitswertes wurde eine Precision-Recall-Curve herangezogen, die einen Schwellenwert berechnet, der den F1 Wert im Training maximiert.

Für das Training des Klassifizierers wurden Konklusionen generiert, auf dessen Daten BART bereits fingetuned wurde - mit dem Ziel, dass sehr semantisch ähnliche Regenerierungen produziert werden und der Klassifizierer dadurch einen hohen Schwellenwert festlegt, um ein Argument als zulänglich zu klassifizieren. Im Testing wird dem Modell deshalb eine gute Generalisierbarkeit abverlangt, das heißt schlecht produzierte Konklusionen, die zusammen mit der originalen Konklusion einen geringen reskalierten Kosinuswert aufweisen, sollten folglich unzulängliche Argumente darstellen. Dies würde gleichzeitig jedoch voraussetzen, dass BART die Trainingskonklusionen inhaltlich beziehungsweise semantisch in der Tat sehr genau reproduziert und über gewisse allgemeine argumentative Inferenzfähigkeiten verfügt, die stark von den Pre-Training und Finetuning Daten abhängig sind. Für ein neuronales Netz erscheint der *Argument Annotated Essays* jedoch zu klein und inhaltlich stark divergierende Konklusionen wiesen dennoch eine hohe semantische Ähnlichkeit auf, die eine Klassenunterscheidung erschwerte.

4. Ergebnisse

Ansatz		**Accuracy**	**Macro Precision**	**Macro Recall**	**Macro F1**
Human upper bound		0.911	0.873	0.903	0.883
Gurcke et al.		0.889	0.882	0.880	0.876
Mein Ansatz	**Durchschnitt**	**0.485**	**0.341**	**0.5**	**0.38**
	Fold 1	0.371	0.186	0.5	0.271
	Fold 2	0.369	0.682	0.504	0.58
	Fold 3	0.646	0.323	0.5	0.392
	Fold 4	0.721	0.360	0.5	0.418
	Fold 5	0.318	0.159	0.5	0.241

Tabelle 1: Ergebnisse des Experimentes (Accuracy, Macro Precision/Recall/F1) im Vergleich zur human upper bound und Gurcke et al.

Die Ergebnisse meines Ansatzes zeigen bei allen Metriken eine deutlich schlechtere Performanz im Vergleich zu der human upper bound und dem Ansatz von Gurcke et al. Bei genauerer Betrachtung der Ergebnisse stellte sich heraus, dass entweder ein sehr hoher Kosinuswert als Schwellenwert festgelegt wurde und dadurch fast alle Konklusionen als unzulänglich klassifiziert

wurden - das sich vor allem in den accuracy Werten von Fold 1, 2 und 5 widerspiegeln (ca. 34% der Daten bestehen aus unzulänglichen Argumenten) - oder ein mittlerer Schwellenwert bestimmt wurde. Bei letzterem wurden fast alle Konklusionen als zulänglich bewertet, was sich in den accuracy Werten von Fold 3 und 4 zeigte, denn circa zwei Drittel der Daten sind als zulänglich annotiert. Die allgemeine Performanz ist deshalb maximal nur gleichauf mit einem Dummy-Klassifizierer, der alle Konklusionen anhand der in den Daten prominentesten Klasse beurteilt.

5. Fazit

Mithilfe des von mir gewählten Ansatzes konnte die Klassifizierungsentscheidung transparenter gemacht werden, jedoch konnte damit weder Evidenz für die von Gurcke et al. aufgestellte Hypothese gefunden noch eine höhere Performanz als ein Dummy-Klassifizierer erreicht werden. Das Merkmal der semantischen Ähnlichkeit hat sich mit diesem Ansatz als nicht repräsentativ herausgestellt und kann den Inhalt einer Konklusion nicht korrekt abbilden. Inhaltlich stark abweichende Konklusionen weisten trotzdem hohe Kosinuswerte auf, weil sich bestimmte Wörter überschneideten oder semantisch ähnliche Wörter enthalten waren. Dadurch konnte kein distinguierender Schwellenwert festgelegt werden, der die beiden Klassen zuverlässig voneinander trennen konnte.

Mögliche Gründe für die schlechten Ergebnisse liegen in der Korpusgröße, die für ein neuronales Netz deutlich zu klein ist, allgemein sehr subjektive Betrachtung und Annotation der *local sufficiency* sowie qualitativ sehr unterschiedliche Reproduktionen der Konklusionen. Das Modell war teils in der Lage, sie inhaltlich gut zu regenerieren, produzierte aber auch sehr abweichende oder sogar qualitativ verbesserte Konklusionen. Es ist ohne Weiteres auch nicht beurteilbar, ob die starken Abweichungen an der Qualität der Prämissen liegen, oder ob das verwendete Modell - auch mit besserem Finetuning - allgemein nicht gut genug bei dieser Aufgabe performieren kann. Sehr gute allgemeine Inferenzfähigkeiten müssten für die Hypothese gegeben sein, damit man anhand der Reproduktionen Rückschlüsse auf die Argumentzulänglichkeit ziehen könnte.

Literatur

[1] Timon Gurcke. *Sufficiency Assessment Generation*. https://github.com/timongurcke/sufficiency-assessment-generation. GitHub Repository. Letzter Aufruf: 08.02.2023. 2021.

[2] Timon Gurcke, Milad Alshomary und Henning Wachsmuth. "Assessing the Sufficiency of Arguments through Conclusion Generation". In: *Proceedings of the 8th Workshop on Argument Mining*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, S. 67–77. DOI: 10.18653/v1/2021.argmining-1.7. URL: https://aclanthology.org/2021.argmining-1.7.

[3] Christian Stab und Iryna Gurevych. *Argument Annotated Essays (version 2)*. 2017. URL: https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2422.

[4] Henning Wachsmuth u. a. "Computational Argumentation Quality Assessment in Natural Language". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, S. 176–187. URL: https://aclanthology.org/E17-1017.