Anwenderforum für Fernmeldetechnik, AFCEA Computer, Elektronik und Automatisierung

AFCEA Bonn e.V. Studienpreis 2021/2022

Kernthesen der Arbeit

Titel der Arbeit:	Multimodal Transformers for Biomedical Text and Knowledge Graph Data
Tag der Einreichung:	3. September 2021
Hochschule:	H-BRS
Name des Verfassers:	Helena Balabin
Betreuender Professor:	Prof. Dr. Paul G. Plöger

*Kurze Beschreibung (1 Seite !) der **Kernthesen**. Was ist die Quintessenz der Arbeit?*

Jüngste Meilensteine aus dem Bereich des Maschinellen Lernens (ML) schlagen sich zunehmend auch bei der Verarbeitung von natürlicher Sprache (NLP) nieder. Stellvertretend seien genannt die Arbeiten von T.Mikolov et al., word2vec [2013], A.Vaswani et al., transformers and attention [2017] und J.Devlin et al., BERT-based embeddings [2018]. Die Güte von Sprachmodellen (Language Models, LMs) und ihre Praktikabilität haben sich damit beträchtlich erweitert. LMs funktionieren dann besonders gut, wenn sie sich auf ein spezielles Fachgebiet beziehen. Große medizinische Textbestände -wie PubMed.gov- können mit rein NLP-basierten Methoden schon mit einer befriedigenden Trefferrate klassifiziert werden. Die vorliegende Masterarbeit wendet LMs exemplarisch auf bio-medizinische Fragen an, ist aber methodisch nicht darauf beschränkt.

Eine schwierige -und auch häufige- Fragestellung könnte in diesem Kontext lauten: angenommen, eine Substanz wurde als wirksam nachgewiesen, für welche Spezies gilt dann diese Aussage? Eine korrekte Antwort müsste Kontextinformationen einbeziehen, d.h. das LM bettet nicht nur den einen betreffenden Satz in seine NLP Darstellung ab, sondern "versteht", dass das ganze Dokument sich auf Mäuse bezieht, aber eben nicht auf Menschen.

Ziel der Arbeit ist, für diese Art von mehr kontext-gebundenen Fragen ein LM so zu erweiterten, dass sich die Trefferraten merklich erhöhen. Gesucht wird eine Integration mit einer zweiten Wissensdarstellung, einem sogenannten Knowlegde Graph (KG). KGn bieten mehr kontextbezogene, relationale Repräsentationen mit Konzepten als Knoten und Relationen als Kanten. Die Kernidee der Integration führt dann auf das Problem, dass die attention-basierten LMs auf bi-partiten Graphen fußen, während KGs ganz allgemeine Graphen sind. Die Lösung gelingt dadurch, dass der KG mit wiederholten Random Walks "abgetastet" wird und als eine Schar von zufälligen beschränkten Wanderungen über den KG dargestellt wird. Dadurch erscheint der KG lokal wie eine Folge von "Sätzen", die man dann mit den entsprechenden NLP Techniken weiterverarbeiten kann. Das gemeinsame Substrat des LM kann dabei als das kartesische Produkt von der LM Darstellung gemeinsam mit der abgetasteten KG-Repräsentation verstanden werden.

Die gleichzeitige Verwendung von beiden Wissensquellen erfordert aber auch ein synchrones Lernen, also eine explizite Assoziation von Textteilen mit den entsprechenden KG Entitäten. die Arbeit schlägt daher den "Sophisticated Transformer" (STonKGs) vor, der auf einer gemeinsamen Darstellung von biomedizinischen LMs und Wissensgraphen KGs trainiert wird. STonKGs entwirft, implementiert und testet eine multimodale Architektur, die auf einem Kreuzkodierer basiert, der die NLP-Attention Technik auf eine Verkettung von Eingabesequenzen aus Text ***und*** KG Tripeln anwendet. STonKGs wird in einem nicht-überwachten Training zunächst angelernt mit über 13 Millionen PubMed Text-Tripel-Paaren, die vom Reasoning Assembler INDRA aus Harvard assembliert wurden. Danach wird per Transfer-learning das LM aufgabenspezifisch nach-trainiert und empirisch validiert mit einer achtfach untergliederten Aufgabenstellung. Diese wird verglichen mit einer reinen LM-basierten und einer rein KG-basierten Baseline. Als Ergebnis ergibt sich, dass insbesondere bei kleineren Datenbeständen und für Aufgaben mit einer größeren Anzahl von Klassen STonKGs zu einer erheblichen Leistungssteigerung führt. Es übertraf im Best Case den F1-Score der besten Baselines um über 15% bei der Aufgabe „Context Annotation“. Sowohl der Quellcode als auch die Testdaten sind öffentlich zugänglich (github.com/ stonkgs/ stonkgs); die Methode kann also ohne weiteres auf weitere biomedizinische Anwendungen verallgemeinert werden.

Die Arbeit ist m.E. preiswürdig aus einer ganzen Reihe von Gründen. So wird der Stand der Forschung mindestens erreicht, wenn nicht sogar überschritten. Dies wird unterstrichen durch eine Einreichung als Erst-Autor zur Publikation beim „Journal of Bioinformatics“, sie befindet sich in der 2.-ten Begutachtungsphase. Das Thema ist hoch-aktuell in einem sich rasend schnell entwickelnden Gebiet. Trotzdem schafft die Arbeit - Abgabedatum 3.9.2021 - Quellen aus April 2021 (!) zu berücksichtigen. Die exzellente Bibliografie ist 15 Seiten lang mit 136 einschlägigen Referenzen. Das Thema ist praxis-relevant und fußt auf einer top-aktuellen Kollaboration von Harvard Medical School mit Fraunhofer SCAI, also zwei weltweit ausgewiesenen Key-playern im Bereich der wissensbasierten Bioinformatik. Die Arbeit implementiert eine innovative Architektur mit sehr überzeugenden Resultaten (Tab. 10). Herausragend gelungen sind die Darstellungen von: 2.2 NLP, 2.3 Network Representation Learning, 2.4 Multimodal Transformers, 3.1 Data Preprocessing and Evaluation Strategy, 3.3 Ablation Studies und die fundierte Diskussion über die Grenzen der Methode.