Erweiterung des HunFlair-Frameworks zur Erkennung genetischer Varianten in biomedizinischen Texten

Kay Steinbauer
Humboldt Universität zu Berlin
April 2023

1. Einleitung

Genetische Veränderungen und Mutationen sind alltäglich, in den meisten Fällen harmlos und wichtig für die Anpassung der Organismen an die Umweltbedingungen, in denen sie leben. Genetische Veränderungen können auch negative Folgen haben, wie z.B. das Down-Syndrom, und sind die Hauptursache für Krebserkrankungen [HAN00]. Der Artikel von Hanahan et al. [HAN00] diskutiert den Zusammenhang zwischen Krebserkrankungen und genetischen Veränderungen: „Several lines of evidence indicate that tumorigenesis in humans […] reflect genetic alterations that drive the progressive transformation of normal human cells into highly malignant derivatives“.

Durch die rasanten Fortschritte in der Medizin, insbesondere in der Onkologie, sind wir heute in der Lage, auf der Grundlage des genetischen Profils eines Menschen individualisierte Behandlungsmethoden anzubieten, um Krankheiten wirksamer behandeln zu können [MAL20]. Das Verständnis der Zusammenhänge zwischen genetischen Varianten und Krankheiten, ist daher für die Medizin von großer Bedeutung.

Wie in [LOU16] beschrieben, nimmt die Datenmenge in der biomedizinischen Informatik exponentiell zu. Auf PubMed, einer Suchmaschine für biomedizinische Forschung, werden über 27 Millionen Artikel bereitgestellt [PUBMED20]. Im Februar 2023 konnten auf PubMed 285.188 Abstracts zur Suchanfrage „genetic variants“ gefunden werden.

Genetische Varianten mittels traditioneller schlagwortbasierter Suche aus wissenschaftlichen Artikeln zu filtern, erweist sich als nicht praktikabel, da viele Autoren sich nicht an standardisierte Nomenklaturen halten und genetische Variationen in natürlicher Sprache umschreiben. Wei et al. [WEI13] geben an: „most mutations are not described in accordance with standard nomenclature (<25% in our corpus)“.

Um diese Daten der medizinischen Forschung zur Verfügung zu stellen, sind daher Modelle zur automatischen Identifikation erforderlich.
Ein häufig verwendeter Ansatz, um diese Varianten effizienter herauszufiltern zu können, ist die Verwendung von Modellen aus dem maschinellen Lernen. tmVar3.0 ist dafür ein State-of-the-Art Beispiel und ist „a machine learning-based approach to optimally recognize variant components“ [WEI22].

2. Grundlagen

Die Identifizierung von genetischen Varianten in biomedizinischen Texten ist ein Sequence Labeling Problem, bei dem einer Folge von Wörtern BIO-Labels zugeordnet werden, um zu kennzeichnen, ob es sich bei einem Wort um den Anfang einer genetischen Variante (B), ein Wort in einer genetischen Variante (I) oder um kein Wort einer genetischen Variante (O) handelt. Eine spezielle Form des Sequence Labeling Problems, mit dem wir uns in dieser Arbeit beschäftigen werden, ist die Named Entity Recognition (NER). NER ist ein Verfahren zur Identifizierung und Klassifizierung bestimmter Entitäten in einem Text [JLI22]. Entitäten können dabei Personennamen, Ortsnamen oder andere Daten sein. Eine spezielle Form der NER, die auf biomedizinische Texte spezialisiert ist, ist die Biomedical Named Entity Recognition (BioNER).
HunFlair [WEB21] ist ein BioNER-Modell, das die Entitäten Zelllinien, Chemikalien, Krankheiten, Gene und Spezies aus biomedizinischen Texten herausfiltern kann. HunFlair setzt dafür eine Bidirektionale LSTM-CRF (Long Short-Term Memory - Conditional Random Field) ein.
Ein LSTM [HOC97] ist ein spezieller Typ eines rekurrenten neuronalen Netzes, das darauf spezialisiert ist, Abhängigkeiten auch über eine längere Sequenz im „Gedächtnis“ zu behalten.
Die LSTM verarbeitet die Eingabe sequentiell von links nach rechts und kann daher nur Abhängigkeiten von Elementen kodieren, die früher in der Sequenz vorkamen. Um auch Abhängigkeiten von zukünftigen Elementen für die Vorhersage eines Labels zu berücksichtigen, wird eine weitere Schicht der LSTM hinzugefügt, die den Input von rechts nach links verarbeitet. Diese Variante eines LSTM, bei der die Eingabe von beiden Enden beginnend verarbeitet wird, wird auch als bidirektionales LSTM bezeichnet [HUA15].
Ein CRF [LAF01] ist ein Modell zur Modellierung von Abhängigkeiten in sequenziellen Daten.
HunFlair erzielt bei der Annotation biomedizinischer Entitäten ähnliche oder sogar bessere Ergebnisse als andere BioNER-Modelle [WEB21] und ist frei über https://github.com/flairNLP/flair verfügbar.

3. Verwandte Arbeiten

Modelle, die genetische Varianten aus biomedizinischen Texten annotieren können, existieren bereits. Ein Beispiel ist SETH [THO16], ein Open-Source-Tool, das 2016 veröffentlicht wurde. Es ist in der Lage, die von Dunnen und Antonarakis [DEN00] und dem Ad Hoc Committee on Mutation Nomenclature [ADH96] beschriebenen Nomenklaturen für menschliche Mutationen und häufige Abweichungen davon zu erkennen. SETH erreicht dabei gute Werte auf öffentlich verfügbaren Korpora. Beispielsweise erreicht SETH eine Precision von 98% und einen Recall von 83% auf dem Korpus von Caporaso et al. [THO16].
Ein weiteres Tool wurde 2022 mit tmVar 3.0 veröffentlicht, das auf den Korpora tmVar [WEI13], OSIRIS [BON06] und Thomas [THO16] in Precision und Recall um durchschnittlich 5 Prozentpunkte besser abschneidet als SETH. SETH und tmVar sind beide frei verfügbare Tools und können unter http://rockt.github.io/SETH/ bzw. https://github.com/ncbi/tmVar3 gefunden werden.
Abbildung 1 zeigt die genetischen Varianten, die von tmVar 3.0 [WEI22], tmVar 2.0 [WEI13] und SETH [THO16] erkannt werden.

Type	Example	tmVar 3.0	tmVar2.0	SETH
SNP	Rs763780	✓	✓	✓
DNA mutation	c.1976A>T	✓	✓	✓
DNA allele	1976A	✓		
DNA change	A>T	✓	✓	✓
Protein mutation	p.Gln659Leu	✓	✓	✓
Protein allele	glutamine at codon 659	✓		
Protein change	methionine to threonine	✓	✓	✓
Other mutations	306 base pair insertion	✓		
Copy number variant	Chr15: 31 833 000–37 477 000 bp deletion	✓		
RefSeq	NM_203475.1	✓		
Chromosome	10q11.12	✓		
Genomic region	Chr10: 46 123 781–51 028 772	✓		✓

Abbildung 1: Genetischen Varianten die von tmVar 3.0, tmVar 2.0 und SETH erkannt werden [WEI22]

4. Ziel

Ziel dieser Arbeit ist es, HunFlair um die Funktionalität der Erkennung von genetischen Varianten zu erweitern. Dabei beschränke ich mich auf die in Abbildung 1 aufgeführten genetischen Varianten. Die Erkennung wird durch ein BiLSTM-CRF (Bidirektionales LSTM-CRF) realisiert werden. Beispielsweise sollte das Modell für die Eingabe *„Ser326Cys polymorphism and risk...“* das Token 'Ser326Cys' als B-Variante kennzeichnen und alle anderen Token in der Eingabe sollten als O (keine Entität) klassifiziert werden.
Dafür wird das BiLSTM-CRF auf dem annotierten Korpora tmVar [WEI13] trainiert. Darüber hinaus soll die Performance der implementierten Funktionalität mit den NER Tools tmVar 3.0, tmVar 2.0 und SETH verglichen werden.

5. Vorgehensweise

Die Konstruktion des NER-Taggers kann in mehrere Schritte unterteilt werden. Dabei lassen sich die Implementierungen der bereits in Hunflair integrierten NER-Tagger als Orientierung verwenden. Die grobe Vorgehensweise ist in Abbildung 2 dargestellt. Zusammengefasst wird eine Auswahl von Korpora erstellt, die alle in einem großen einheitlichen Schema vereint werden, mit dem das maschinelle Lernmodell trainiert wird. Am Ende wird das Modell in der Lage sein, Vorhersagen auf ungesehenen Text zu treffen.

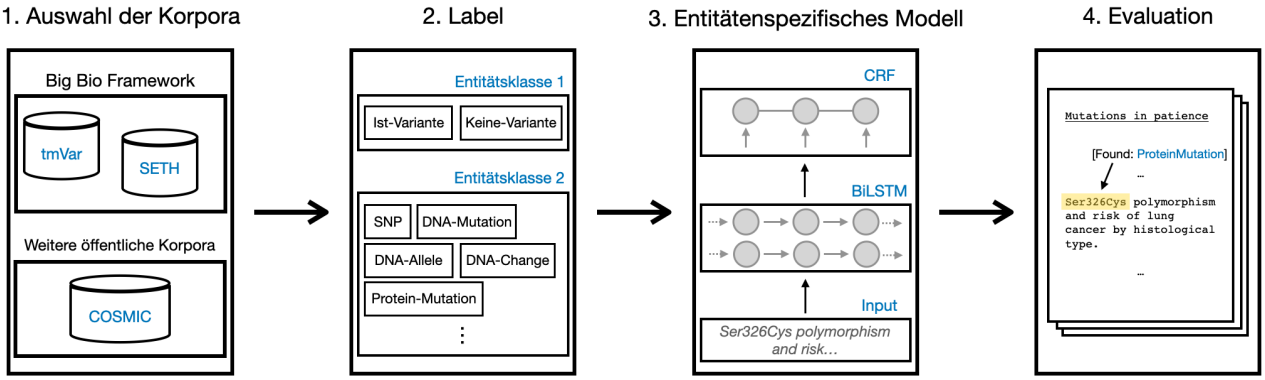

Abbildung 2: Überblick über die Konstruktion des NER-Taggers

5.1 Auswahl der Korpora

Eine wichtige Entscheidung wird die Auswahl der Trainings- und Testdaten sein, da diese entscheidend dafür sind, wie gut das System letztendlich genetische Varianten erkennen kann. Das Modell mit dem tmVar3-Korpus trainiert, der auch vom State-of-the-Art-Tool tmVar 3.0 genutzt wird. Darüber hinaus wird es auch auf den Korpora Cosmic [COSMIC] und BioRed [BIORED] trainiert, um eine breitere Datengrundlage zu erhalten und ähnliche Ergebnisse bei der Erkennung von genetischen Varianten zu erzielen. Die Integration der Datensätze erfolgt über das Github-Framework BigBio [BIGBIO], das bereits eine Vielzahl von Variantendatensätzen enthält und diese in einem einheitlichen Format zur Verfügung stellt, das die Weiterverarbeitung erleichtert.

5.2 Label

Die Labels bezeichnen die Entitätsklassen, die vom NER-Tagger vorhergesagt werden. Zur Erkennung wird dabei das BIO Tagging-Schema verwendet. Für die Wahl der Labels gibt es zwei mögliche Ansätze für einen NER-Tagger zur Erkennung von genetischen Varianten mit unterschiedlichem Granularitätsgrad. Eine mögliche Option ist, dass der NER-Tagger den Input in die Kategorien

B-Variante, I-Variante und O aufteilt und in einem späteren Schritt die genaue Art der genetischen Variante identifiziert wird. Eine andere Möglichkeit wäre, dass der NER-Tagger den Input direkt nach den spezifischen Variantentypen aus Abbildung 1 annotiert, um in einem Schritt sowohl das Vorhandensein als auch den spezifischen Typ der genetischen Variante zu identifizieren.
Ein Beispiel, das die Unterschiede der Labeling-Ansätze verdeutlicht, ist die Verarbeitung des Eingabesatzes "Ser326Cys polymorphism and risk…“.
Beim ersten Beispiel wird Ser326Cys nur als B-Variante gekennzeichnet, bei der zweiten Variante wird Ser326Cys als B-ProteinMutation.

5.3 Entitätenspezifisches Modell

Zur Konstruktion des NER-Taggers wird ein BiLSTM-CRF verwendet.
Das BiLSTM-CRF in HunFlair basiert auf dem Flair Language Model, das auf 3 Millionen PubMed-Artikeln und 25 Millionen Abstracts trainiert wurde. Das Model wird mit einem LSTM mit einer Hidden-Layer-Größe von 2048 und einer Batch-Größe von 256 trainiert [SUP20]. Als Word Embedding wird fastText [BOJ17] verwendet. Die Word Embeddings werden dann als Input für die BiLSTM verwendet, die sie weiter verarbeitet und dann als Input an die CRF-Schicht weitergibt. Die finale Ausgabe wird dann mit dem Viterbi-Algorithmus aus der CRF-Schicht berechnet. Die technischen Details lassen sich im Paper von Weber et al., 2020 [WEB20] nachlesen.

5.4 Evaluation

Um die Performance der implementierten Funktionalität zu bewerten, wird ein In-Corpus und Cross-Corpus-Evaluationsansatz gewählt. Bei der In-Corpus Evaluation wird das Modell auf einer Teilmenge eines Corpus trainiert und anschließend auf einer anderen Teilmenge desselben Corpus getestet. In der Arbeit von Mita et al. [MIT19] wird aufgezeigt, dass „empirical evaluation results […] considerably vary depending on the corpus, suggesting that a single-corpus evaluation can be unreliable. Therefore, cross-corpora evaluation should be applied“. Bei der Cross-Corpus Evaluation wird das Modell auf anderen Korpora getestet als es trainiert wurde.
Für die Cross-Corpus Evaluation werden zusätzlich die Korpora OSIRIS [BON06] und Thomas [THO16] verwendet. Ziel der Kombination der beiden Bewertungsansätze ist die Gewinnung eines möglichst genauen Überblicks über die Performance. Die Ergebnisse werden dann mit den State-of-the-Art Tools SETH, tmVar 3.0 und tmVar 2.0 hinsichtlich der Metriken Precision, Recall und F1-Score verglichen werden. Die Performance dieser Tools auf den genannten Korpora ist in Abbildung 3 dargestellt.

Corpus	Task	Method	Precision (%)	Recall (%)	F-score (%)
tmVar	NER	tmVar 3.0	94.01	88.86	91.36
		tmVar 2.0	98.22	80.64	88.57
		SETH	97.92	68.77	80.79
	Normalization	tmVar 3.0	96.99	91.71	94.28
		tmVar 2.0	94.49	77.25	85.00
		SETH	86.51	69.91	77.33
OSIRIS	NER	tmVar 3.0	98.62	84.98	91.30
		tmVar 2.0	99.53	83.00	90.52
		SETH	96.43	74.70	84.19
	Normalization	tmVar 3.0	97.72	84.58	90.68
		tmVar 2.0	97.20	80.62	88.14
		SETH	94.21	69.38	79.91
Thomas	NER	tmVar 3.0	92.26	91.30	91.78
		tmVar 2.0	82.46	97.04	89.16
		SETH	84.43	69.39	76.18
	Normalization	tmVar 3.0	91.01	90.32	90.67
		tmVar 2.0	89.94	88.24	89.08
		SETH	95.58	57.50	71.80

Abbildung 3: Performance von SETH, tmVar 3.0, tmVar 2.0 auf den Korpora tmVar, OSIRIS und Thomas [WEI22]

6. Weiterführende Arbeiten

Das Erkennen von genetischen Variationen in Texten ist nur der erste Schritt eines BioNER-Tools. Um das Ganze praktisch nutzbar zu machen, müssen die gefundenen Entitäten normalisiert werden, denn erst dann steht die mit einer genetischen Variation verbundene Information zur Verfügung. Derzeit ist HunFlair nicht in der Lage die gefundenen Entitäten zu normalisieren.

Im Normalisierungsschritt werden die Entitäten mit Einträgen in biomedizinischen Datenbanken verglichen und bei Übereinstimmung gematcht. Dafür muss ein Modul konstruiert werden, das in der Lage ist, die in natürlicher Sprache beschriebenen genetischen Variationen auf Datenbankbezeichner abzubilden. Dabei existiert keine zentrale Datenbank, die alle genetischen Variationen umfasst. Es wird dafür eine Vielzahl von Datenbanken durchsucht werden müssen, die jeweils Millionen von Einträgen enthalten können.

Das Paper von Thomas et. al [THO11] befasst sich unter anderem mit den Schwierigkeiten des Normalisierungsprozesses.

7. Literatur

[BIGBIO] Big Bio Framework: https://github.com/bigscience-workshop/biomedical

[BIORED] BioRED: https://github.com/bigscience-workshop/biomedical/tree/main/bigbio/hub/hub_repos/biored

[BOJ17] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, "Enriching Word Vectors with Subword Information" in Transactions of the Association for Computational Linguistics, vol. 5, pp. 135-146, 2017.

[BON06] J. Bonis, L. I. Furlong, and F. Sanz, "OSIRIS: a tool for retrieving literature about sequence variants" in Bioinformatics, vol. 22, no. 20, pp. 2567-2569, Oct. 2006

[COSMIC] COSMIC: https://cancer.sanger.ac.uk/cosmic

[DEN00] J. T. den Dunnen and S. E. Antonarakis, "Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion" in Human Mutation, vol. 15, no. 1, pp. 7-12, Jan. 2000

[HAN00] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer" in Cell, vol. 100, no. 1, pp. 57-70, Jan. 2000

[HOC97] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory" in Neural Computation, vol. 9, no. 8, pp. 1735-1780, Nov. 1997

[HUA15] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging" in arXiv:1508.01991 [cs.CL], Aug. 2015

[JLI22] J. Li, A. Sun, J. Han and C. Li, "A Survey on Deep Learning for Named Entity Recognition," in IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 1, pp. 50-70, Jan. 2022

[LAF01] J. D. Lafferty, A. McCallum and F. C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data" in Proceedings of the Eighteenth International Conference on Machine Learning, San Francisco, CA, USA, 2001, pp. 282-289, Morgan Kaufmann Publishers Inc., ISBN: 1-55860-778-1.

[LOU16] J. Luo, M. Wu, D. Gopukumar, and Y. Zhao, "Big Data Application in Biomedical Research and Health Care: A Literature Review“ in Biomedical informatics insights, vol. 8, pp. 1-10, Jan. 2016

[MAL20] E. R. Malone, M. Oliva, P. J. B. Sabatini, et al., "Molecular profiling for precision cancer therapies" in Genome Medicine, vol. 12, no. 1, p. 8, 2020

[MIT19] M. Mita, T. Mizumoto, M. Kaneko, R. Nagata, and K. Inui, "Cross-Corpora Evaluation and Analysis of Grammatical Error Correction Models — Is Single-Corpus Evaluation Enough?" in arXiv:1904.02927 [cs.CL], 2019

[PUBMED20] https://pubmed.ncbi.nlm.nih.gov/

[THO11] P. E. Thomas, R. Klinger, L. I. Furlong et al., "Challenges in the association of human single nucleotide polymorphism mentions with unique database identifiers" BMC Bioinformatics, vol. 12, no. Suppl 4, p. S4, 2011

[THO16] P. Thomas, T. Rocktäschel, J. Hakenberg, Y. Lichtblau, and U. Leser, "SETH detects and normalizes genetic variants in text“ in Bioinformatics, vol. 32, no. 18, pp. 2883-2885, Sep. 2016

[WEB20] L. Weber, J. Münchmeyer, T. Rocktäschel, M. Habibi and U. Leser, "HUNER: improving biomedical NER with pretraining" in Bioinformatics, vol. 36, no. 1, pp. 295-302, 2020

[WEB21] L. Weber, M. Sänger, J. Münchmeyer, M. Habibi, U. Leser, and A. Akbik, "HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition“ in Bioinformatics, vol. 37, no. 17, pp. 2792-2794, Sep. 2021

[WEI13] C. H. Wei, B. R. Harris, H. Y. Kao, Z. Lu, "tmVar: a text mining approach for extracting sequence variants in biomedical literature“ in Bioinformatics, vol. 29, no. 11, pp. 1433-1439, June 2013

[WEI22] C. H. Wei, A. Allot, K. Riehle, A. Milosavljevic, and Z. Lu, "tmVar 3.0: an improved variant concept recognition and normalization tool" in Bioinformatics, vol. 38, no. 18, pp. 4449-4451, Sep. 2022