Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik

Wörterbuch-basierte Normalisierung biologischer Entitäten

Exposé zur Bachelorarbeit

eingereicht von: Ronja Spiegelberg
geboren am: 03.07.1998
geboren in: Frankfurt (Oder)

Gutachter/innen: Prof. Dr. Ulf Leser
Dr. rer. nat. Manuela Benary

eingereicht am: .....

1 Exposé

1.1 Einleitung und Motivation

Die Erfassung und Verarbeitung von biologischen Daten fordert verschiedene Bereiche der Medizin und insbesondere der Bioinformatik heraus. Es sind leistungsfähige Systeme und Prozesse notwendig, um eine effiziente Arbeitsweise zu ermöglichen. Eingesetzte und benötigte Anwendungen haben die Aufgabe, gegebene Informationen umfassend aufzubereiten oder zu verarbeiten. Durch diesen Prozess werden die Informationen erst zugänglich oder weiterverarbeitbar. Biomedizinische Daten können in heterogenen, verteilten und sich rapide entwickelnden Quellen oder Versionen vorliegen und auch widersprüchliche Aussagen enthalten. Diese und weitere Besonderheiten sollen möglichst transparent, uniform, vollständig und automatisiert aufbereitet werden, um sie in verschiedenen biomedizinischen Umgebungen einsatzfähig zu machen. [1]

Speziell in der personalisierten Krebsmedizin (oder Präzisionsonkologie, kurz PO) müssen diese Informationen aus einer Vielzahl von unterschiedlichsten Datenquellen entnommen werden, sodass eine möglichst umfassende Informationsbasis und damit präzisere Recherche ermöglicht werden kann. [2] Es werden also Datenintegrationssysteme benötigt, um die wachsende Datenbasis laufend aktualisieren zu können, ohne großen manuellen Aufwand zu verursachen. Eine Reihe von präzisionsonkologischen Wissensbasen (Precision Oncology Knowledge Base, POKBs) entstehen aus Kuratierung von Informationen aus wissenschaftlicher Literatur. POKBs stellen strukturierte Informationen bezüglich der Beziehungen zwischen Genen und Varianten, sowie Medikamenten und weiteren relevanten Verknüpfungen bereit. CIViC [3] (Clinical Interpretations of Variants in Cancer) und OncoKB [4] sind Datenbanken zur Interpretation von Krebsvarianten, um nur zwei populäre Stellvertreter zu nennen. Beide legen Fokus auf Bereitstellung von Informationen über Diagnose und Therapie sowie begünstigende Faktoren und vorbeugende Maßnahmen, dennoch unterscheiden sie sich in Anzahl der enthaltenen Gene und Detailgrad der weiteren Informationen.

Bei der Arbeit mit biomedizinischen Wissen liegen Informationen zunächst textuell in natürlicher Sprache vor (Befunde, Notizen, Paper). Um die Daten verwenden zu können, müssen diese gelesen oder verarbeitet werden. Da dieser Vorgang nicht skaliert, wird sich an Verfahren der Verarbeitung von Natürlicher Sprache (NLP oder spezieller bioNLP) bedient. Typischerweise werden Texte im bioNLP mit Named Entity Recognition (NER) und Named Entity Normalization (NEN) analysiert. NER hat dabei die Aufgabe den Begriff von seiner Start- bis zu seiner Endposition zu erkennen. NEN (auch Grounding oder Normalisierung genannt) bildet diesen Begriff dann auf eine eindeutige Kennzeichnung aus bestimmten Namensräumen ab. [5] Beispielsweise wird das Gen das mit dem Namen TP53 und der ID 7157 (in NCBI Reference Sequences) bei Mutation mit Krebs in Verbindung gebracht. [6] Da für dieses Gen aber viele Synonyme, wie P53, BCC7, LFS1, BMFS5 oder

TRP53 existieren, wird in der Praxis die Nutzung eines Werkzeugs zur Normalisierung notwendig, um Synonyme oder variierende Schreibformen korrekt zuzuordnen.

Eine erstrebenswerte Verbesserung eines bereits bestehenden Systems ist eine Erhöhung der Qualität der erzeugten Daten. Von einer präziseren Arbeitsweise des Programms kann nicht nur der Nutzer selbst, sondern im Resultat auch Forschung und Patient profitieren. Im Kontext dieser Arbeit wird dies bei der Normalisierung im Schritt der Entitätenerkennung mittels einer Senkung der Fehlerrate durch optimiertere Ähnlichkeitsbestimmung angestrebt. Das hierbei zu modifizierende Werkzeug heißt PREON (siehe Kapitel 1.3).

1.2 Zielstellung

Im Rahmen dieser Bachelorarbeit wird eine wörterbuch-basierte Normalisierung für biologische Entitäten entwickelt. Hierbei wird die Berechnung der Editierdistanz (Levenshtein-Distanz, kurz LD) von PREON entsprechend zu einer Gewichteten Editierdistanz (Weighted-Levenshtein-Distanz, kurz WLD) erweitert. Damit können geeignete Konfigurationen der WLD auf den verwendeten Datensätzen gefunden werden, sodass eine qualitativ messbare Verbesserung der Entitäten-Erkennung (bei der Normaisierung) erkennbar wird.

Aufbauend auf die Modifikation PREONs (kurz PREON*) wird ein adäquater Vergleich der untersuchten Tools hervorgebracht, um damit eine Aussage über Fähigkeiten und Vergleichbarkeit zu machen. Hierzu wird auf einer Sammlung von Korpora [2] und einer Reihe von entsprechenden Ontologien [3] evaluiert. Dabei werden typische Metriken (Accuracy, Precision, F1-Score,...) verwendet. Die Ergebnisse von PREON* mit verschiedenen Kosten-Konfigurationen der WLD wird auf den genannten Korpora und Ontologien evaluiert und dabei gewonnene Erkenntnisse herausgearbeitet. Untersucht wird (soweit möglich) auch, wie schnell und präzise PREON* im Vergleich zu GILDA ist.

[2] BC2 (human), BC2 (full), NLM-Gene, NCBI Disease, BC5CDR (d), BC5CDR (c), NLM-Chem, Linnaeus, S800. Korpora beinhalten von Experten annotierte Textsammlungen mit Zurordnung der korrekten Entität zu ID.

[3] CTD Diseases, CTD Chemicals, NCBI Taxonomy, NCBI Gene (human), NCBI Gene, Cellosaurus. Ontologien sind hier als äquivalent zu POKBs zu verstehen.

1.3 Verwandte Arbeiten

Es existieren eine Reihe an Einsatzmöglichkeiten von Distanzmaßen zur Ähnlichkeitsbestimmung. Im Kapitel *Methoden der Linked Data Integration* aus dem Buch *Linked Enterprise Data* [7] wird eine Art Workflow zur Zusammenführung unterschiedlicher Datensätze (Datenintegration) im Einsatz von Linked-Data im Unternehmenskontext beschrieben. Es wird hierbei unter anderem ein Überblick über populäre Strategien zur Ähnlichkeitsbestimmung von verschiedenen Werten vorgestellt und geeignete Anwendungsfälle benannt. Zur Evaluierung werden ebenfalls verschiedene Methoden, wie Genauigkeit, Trefferquote und deren Kombination (F-Measure) vorgestellt. Grundaussage des Kapitels ist, dass Verfahren zur Datenzusammenführung eine wichtige Basis für die zukünftige Forschung und Entwicklung (insbesondere von Linked-Data) bilden.

In dem Artikel *Levenshtein Distance Technique in Dictionary Lookup Methodes: An Improved Approach* [8] werden wörterbuch-basierte Methoden zur Erkennung von Wörtern (der Länge drei und fünf) im Kontext visueller Worterkennung (beispielsweise Scan eines Dokuments digitalisieren) beschrieben. Es wird zwischen der klassischen Levenshtein-Distanz und einer modifizierten Variante, einer speziell gewichteten Levenshtein-Distanz (WLD), verglichen. Dafür wurden Buchstaben in visuell ähnliche Gruppen aufgeteilt, angelehnt an handschriftliche Ähnlichkeit, die dann eine geringere Gewichtung (also mehr Ähnlichkeit aufweisen) erhalten. Die Ergebnisse deuten bei den hier getesteten Datensätzen mit entsprechenden Wortlängen darauf hin, dass die angepasste WLD eine bessere Erkennungsrate aufweist als die normale LD und damit weniger fehleranfällig ist. Im Kontext der Worterkennung erleichtert dieses Verfahren die Nutzerfreundlichkeit und senkt den manuellen Aufwand zur Korrektur.

PREON [9] ist eine in Python implementierte Bibliothek zur Normalisierung biomedizinischer Entitäten in Datenbanken für Krebsarten und Medikamentennamen speziell für Datenintegrationsprojekte. Es basiert auf einem mehrstufigen System aus drei Matching-Algorithmen, wobei nur bei Bedarf auf die komplexeren Algorithmen zurückgegriffen wird, um eine kürzere Laufzeit zu erzielen. Aufgrund der Verwendung von mehreren verschiedenen Datenbanken lassen sich Redundanzen in enthaltenen Informationen kaum vermeiden. Daher ist eine schnelle und akkurate Vorgehensweise bei der Verarbeitung der Einträge notwendig. Da wortweise Datenbankeinträge genutzt werden, PREON sich daher nur auf syntaktische Namensmerkmale stützt, sollte es im Unterschied zu textbasierten NEN Methoden deutlich effizienter normalisieren können.

GILDA [10] - ein ähnliches Werkzeug zur Normalisierung von biomedizinischen Entitäten - wurde zur Normalisierung mittels eines Scoring-basierten Algorithmus entwickelt. Zusätzlich werden hier maschinell erlernte Disambiguierungsmodelle bereitgestellt, die den Kontext des Wortes miteinbeziehen können.

1.4 Vorgehen

Begonnen wird mit der Datenvorverarbeitung der Korpora und Ontologien. Im ersten Schritt werden die Daten so aufbereitet, dass sie sich für die Normalisierung mit PREON und GILDA eignen. Hierfür werden für Experimente überflüssige Informationen ignoriert. Da diese zum Teil in nicht tabellarischer Form vorliegen, werden geeignete Verfahren zur Verarbeitung gewählt. Im nächsten Schritt wird jeweils die Einspeisung der Daten in PREON und GILDA vorbereitet. Das bedeutet jeweils eine Nutzung der gegebenen Schnittstellen.

Zentrales Thema der Arbeit ist die Modifikation der Editierdistanz (LD), hierbei wird PREON um eine Gewichtung bei der Kostenvergabe erweitert. Es wird die bereits vorhandene Implementierung der LD auf eine WLD umgestellt. Durch geschickte Wahl der Kosten sollen Konfigurationen gefunden werden, die weniger falsche Matches im Normalisierungsprozess zulässt. Es werden verschiedene Methoden genutzt um sinnvolle Kosten zu vergeben. Bei verscheidenen Arten von Ursachen der Fehlerentstehung gibt es eine Menge an möglichen Problemlösungen. So eignen sich für typographische Fehler (die vielleicht beim Abtippen der Quelldaten entstanden sind) eher Kosten in Abhängigkeit zu den auf der Tastatur liegenden Nachbar-Tastenabstände. Aber für Fehler die auf handschriftlichen Uneindeutigkeiten oder auf ähnlichem Klang basieren, eignen sich vielleicht Kosten in Abhängigkeit zu sich visuell oder auditiv ähnelnden Buchstaben. Besonders bei biomedizinischen Entitäten entscheidet oft schon ein einziger Buchstabe oder eine einzige Zahl in der Bezeichnung darüber, welcher Domäne sie zugeordnet werden. Auch Falsche Zuordnungen, die durch solche Phänomene entstehen, müssen berücksichtigt werden. Diese und weitere Fehlerquellen werden untersucht mit dem Ziel möglichst gute Kosten für die WLD zu identifizieren.

Für die Evaluierung werden unter Verwendung der bereitgestellten Korpora und Ontologien typische Metriken berechnet und untersucht. Anhand dieser Werte lassen sich Aussagen über allgemeine Genauigkeit und Verbesserung gegenüber PREON und GILDA machen. Gemessene Ergebnisse werden diskutiert und ausführlich präsentiert.

Des Weiteren wird PREON mit GILDA verglichen, sowohl strukturell als auch im Hinblick auf die Laufzeit und verwendeten Metriken (soweit möglich). Strukturelle Vergleichspunkte werden zwischen den zu Grunde liegenden lexikalischen Ressourcen und den Normalisierungsprozessen (Algorithmen, Verfahren) gesetzt. Dabei werden Vor- und Nachteile der jeweils genutzten Verfahren im Kontext der biomedizinischen Einsatzfähigkeit herausgearbeitet.

2 Referenzen

[1] Johannes Starlinger u. a. „Variant information systems for precision oncology“. In: *BMC Medical Informatics and Decision Making* 18.1 (Nov. 2018). DOI: 10.1186/s12911-018-0665-z. URL: https://doi.org/10.1186/s12911-018-0665-z.

[2] Steffen Pallarz u. a. „Comparative Analysis of Public Knowledge Bases for Precision Oncology“. en. In: *JCO Precision Oncology* 3 (2019), S. 1–8. DOI: 10.1200/PO.18.00371. URL: https://doi.org/10.1200/PO.18.00371.

[3] *CIViC - Clinical Interpretation of Variants in Cancer — civicdb.org*. https://civicdb.org/pages/about. [Accessed 2023-Aug-31]. 2022.

[4] *OncoDB — oncodb.org*. https://oncodb.org/index.html. [Accessed 2023-Aug-28]. 2022.

[5] Ming-Siang Huang u. a. „Biomedical named entity recognition and linking datasets: survey and our recent development“. In: *Briefings in Bioinformatics* 21.6 (Juni 2020), S. 2219–2238. DOI: 10.1093/bib/bbaa054. URL: https://doi.org/10.1093/bib/bbaa054.

[6] NCBI Reference Sequences. *7157 - Gene ResultTP53 tumor protein p53 [(human)]*. [Accessed 2022-Nov-22]. 2016. URL: https://www.ncbi.nlm.nih.gov/gene/7157.

[7] Robert Isele. „Methoden der Linked Data Integration“. In: *Linked Enterprise Data*. Springer Berlin Heidelberg, 2014, S. 103–120. DOI: 10.1007/978-3-642-30274-9_5. URL: https://doi.org/10.1007/978-3-642-30274-9_5.

[8] Rishin Haldar und Debajyoti Mukhopadhyay. „Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach“. In: *arXiv.org* (2011). DOI: 10.48550/ARXIV.1101.1232. URL: https://arxiv.org/abs/1101.1232.

[9] Arik Ermshaus u. a. „preon: Fast and accurate entity normalization for drug names and cancer types in precision oncology“. In: *bioRxiv* (2023). DOI: 10.1101/2023.05.22.540912. eprint: https://www.biorxiv.org/content/early/2023/05/22/2023.05.22.540912.full.pdf. URL: https://www.biorxiv.org/content/early/2023/05/22/2023.05.22.540912.

[10] Benjamin M Gyori, Charles Tapley Hoyt und Albert Steppi. „Gilda: biomedical entity text normalization with machine-learned disambiguation as a service“. In: *Bioinformatics Advances* 2.1 (Jan. 2022). Hrsg. von Cecilia Arighi. DOI: 10.1093/bioadv/vbac034. URL: https://doi.org/10.1093/bioadv/vbac034.