Flache und semantische Verarbeitung von Namen biochemischer Verbindungen

Henriette Engelken*[1,2], Martin Golebiewski*[1], Meik Bittkowski[1], Fritz Hamm[2], Jasmin Saric[1,3], Ulrike Wittig[1], Wolfgang Müller[1], Uwe Reyle[2] und Isabel Rojas[1]

[1]EML Research gGmbH, Schloss-Wolfsbrunnenweg 33, 69118 Heidelberg
[2]Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, Azenbergstraße 12, 70174 Stuttgart
[3]jetzt: Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorferstr. 65, 88397 Biberach / Riß
*E-Mail: engelken@eml-r.org, golebiewski@eml-r.org

Abstract

In den Biowissenschaften ist Termverarbeitung für Information Retrieval und Information Extraction, für Data Mining und für die Datenintegration in wissenschaftlichen Datenbanken von großer Bedeutung. Die Erkennung, Identifizierung und chemische Klassifizierung ist insbesondere für Molekülnamen nötig, welche häufig in wissenschaftlichen Publikationen, Datenbanken und Patenten vorkommen und die wesentlich für das Verständnis des Inhalts dieser Dokumente sind. Eine eindeutige Bezeichnung einer chemischen Verbindung ist ihre chemische Struktur. In Publikationen und Datenbanken werden jedoch oft ausschließlich Namen verwendet. Diese weisen Besonderheiten auf, welche ihre automatische Identifizierung und Klassifizierung erschweren. Zu nennen sind v. a. Synonymie, d. h. Bedeutungsgleichheit unterschiedlicher Namen, und Unterspezifikation, d.h. nicht vollständige Bestimmung der Namensbedeutung.

Die Namensidentifizierung kann durch Matching zu einer Referenzliste (Datenbank, Ontologie) erreicht werden. Wir haben ein Programm[1] zum normalisierten Namensmatching entwickelt. Die Regeln zur Normalisierung repräsentieren Expertenwissen und beinhalten u. a. morphosyntaktische Umformungen der Namen – z. B. von Suffixen zu gleichbedeutenden Präfixen (z. B. *-phosphate* zu *phospho-*). Zudem werden synonyme Substrings paarweise ersetzt, welche wir mit einem statistischen Verfahren gewonnen haben. Durch die implementierten Namenstransformationen können synonyme Namen matchen, welche durch exaktes Stringmatching nicht gefunden werden.

Unser zweites System hat zum Ziel ausgehend von einer linguistischen Namensanalyse[2] die Molekülstruktur zu rekonstruieren. Diese ist eindeutig und enthält die chemischen Eigenschaften. Die linguistischen Bausteine (Morpheme) jedes Namens liefern bestimmte Constraints über die von diesem Namen bezeichnete chemische Struktur, woraus wir Constraint Satisfaction Probleme über Graphenvariablen modellieren. Mit Hilfe eines Constraintlösers können dadurch alle bezeichneten chemischen Strukturen, auch für unterspezifizierende und Klassen-Namen, bestimmt werden und in der Folge zum semantischen Matching von synonymen Namen und zur Klassifikation dienen.

---

[1] Web-Interface: `http://sabiork.villa-bosch.de/normaWeb`

[2] vgl. Kremer et al. (2006): Analysing and Classifying Names of Chemical Compounds with CHEMorph. In *Proceedings of the 2nd International Symposium on Semantic Mining in Biomedicine*, Jena, Germany, S.37-43.