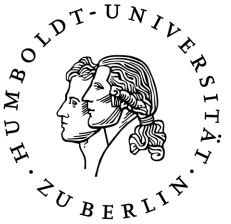
HUMBOLDT-UNIVERSITÄT ZU BERLIN



Exposé zur Studienarbeit

„Chemical Structure Search for Text Corpora“

Autor: Marc N. Bux*

Betreuer: Prof. Dr. Ulf Leser
Josef Scheiber (Novartis AG)

1 Motivation

Forscher der Lebenswissenschaften stehen häufig vor der Aufgabe, aus großen Dokumentmengen wichtige Informationen zu extrahieren ohne dabei versehentlich essentielle Informationen zu übersehen. Banville beschreibt in [Ban06] die Verschiebung der wissenschaftlichen Recherche vom „information gathering“ hin zum „information mining“ als Folge des Überangebots an wissenschaftlichen Publikationen und Erkenntnissen.
Abbildung 1 verdeutlicht wie häufig die Entwicklung von Medikamenten vor allem in späten Phasen abgebrochen werden muss. Hohe Kosten für Forschung und Entwicklung sowie eine kurze Patentzeit bewirken, dass derzeit nur in etwa jedes zehnte Medikament, das auf den Markt kommt, die Entwicklungskosten wieder einspielen kann. Wie Kola und Landis in [KL04] beschreiben ist daher die frühe Erkennung von Problemen, die zum Scheitern des Medikaments in einer späteren Entwicklungsphase führen könnten, von großer Wichtigkeit.
Für die Neu- und Weiterentwicklung von Medikamenten spielen insbesondere Zusammenhänge zwischen chemischen Verbindungen und ihren biologischen bzw. chemischen Auswirkungen eine wichtige Rolle. Text Mining Applikationen, die bei der Erkennung solcher Zusammenhänge behilflich sind, können daher für die pharmazeutische Forschung von großem Nutzen sein.
Beispielsweise könnte eine Suche nach Dokumenten, in denen eine Substruktur der chemischen Struktur von Aspirin (2-Acetoxybenzoesäure) sowie das Wort „cancer“ vorkommt Hinweise auf mögliche unbekannte Nebenwirkungen des Medikaments geben. Es sind viele ähnliche Szenarien vorstellbar, in denen die Verbindung von Keyword-Suche und chemischer Strukturssuche in Text helfen könnte, essentielle Zusammenhänge zu erkennen.

*bux@informatik.hu-berlin.de

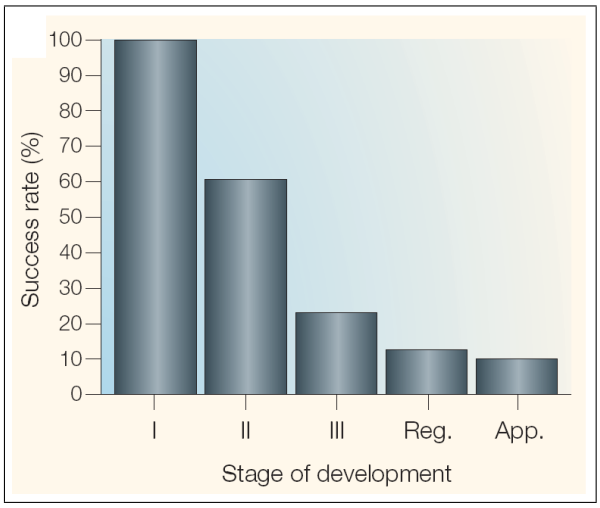


Abbildung 1: Rate, mit der Wirkstoffe einer Entwicklungsphase in die nächste Entwicklungsphase übergehen (Reg. - Registration, App. - Approval) [KL04]

Eine solche Kombination aus Keyword-Suche und Suche chemischer Strukturen existiert derzeit jedoch noch nicht. Das liegt vor allem daran, dass es zwar bereits Applikationen für die Suche nach Strukturen in chemischen Datenbanken wie *Chemspider*[1] und *PubChem*[2] gibt, nicht jedoch für die Suche in Text. Aufgrund der großen Menge von Publikationen in den Lebenswissenschaften erfordert eine Strukturseuche in Text mit akzeptabler Laufzeit, dass zuvor ein Index chemischer Strukturen für den zu durchsuchenden Text erstellt wurde.

2 Zielstellung

Ziel dieser Studienarbeit ist die Entwicklung einer Applikation zur Erkennung chemischer Strukturen in Text und Indexierung dieser Strukturen in einer Datenbank. Diese Datenbank ermöglicht dann eine schnelle Suche chemischer Strukturen sowie ihrer Sub- und Superstrukturen in dem indexierten Text. Die Suche innerhalb einer vorab erstellten Datenbank ist aufgrund der deutlich geringeren Laufzeit einer Online-Suche in Text vorzuziehen. Zum Erstellen der Datenbank sind im Wesentlichen drei Schritte erforderlich:

1. Erkennung von Bezeichnungen für chemische Strukturen in Text mit Hilfe einer externen Applikation mit guter Performance und niedriger Laufzeit
2. Umwandlung der Bezeichnungen in ein eindeutiges maschinenlesbares Format (*InChI*)
3. Registrierung der gefundenen chemischen Strukturen in einer Datenbank

Eine schnelle Suche nach chemischen Strukturen könnte in Kombination mit einer Keyword-Suche nicht nur große Mengen von Text auf überschaubare Dimensionen reduzieren, sondern möglicherweise auch dabei helfen, wertvolle Erkenntnisse aus großen Textmengen zu gewinnen. Eine schematische Darstellung über den Ablauf einer solchen kombinierten Suche ist in Abbildung 2 zu sehen.

[1] http://www.chemspider.com/StructureSearch.aspx
[2] http://pubchem.ncbi.nlm.nih.gov/search/search.cgi

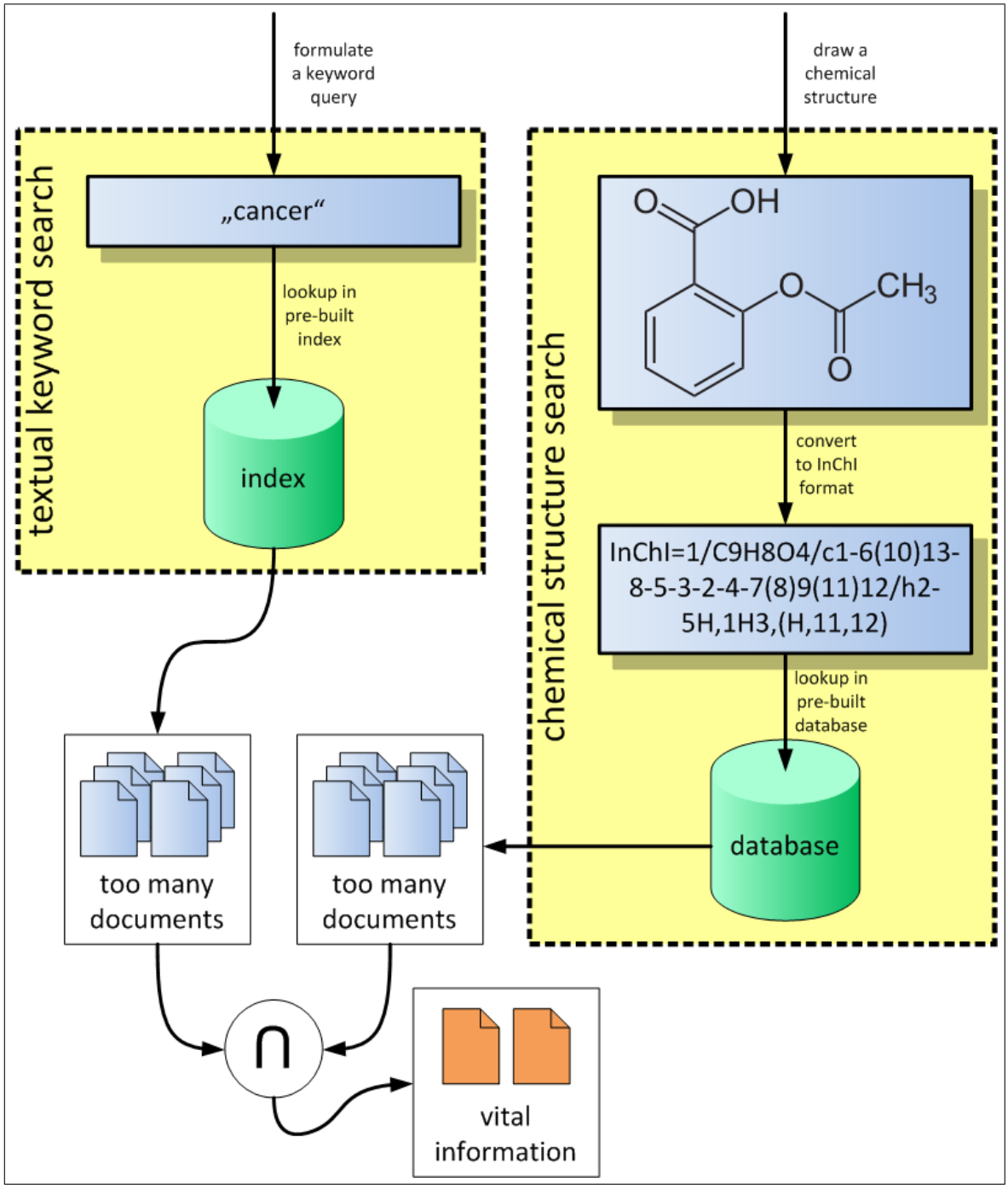


Abbildung 2: Exemplarischer Ablauf einer kombinierten Keyword- und Struktursuche

3 Herangehensweise

3.1 Extraktion chemischer Entitäten

Die Lokalisierung von Bezeichnungen chemischer Strukturen in Text Korpora bringt zwei wesentliche Herausforderungen mit sich. Zum Einen existieren sehr vielfältige Methoden zur Benennung chemischer Strukturen (siehe Tabelle 1), was die automatische Erkennung chemischer Entitäten erschwert. Ein einheitlicher und eindeutiger Standard ist zwar vorhanden, bislang aber nicht gebräuchlich. Selbst gängige Standards wie der *IUPAC*-Standard (*International Union of Pure and Applied Chemistry*) erlauben mehrere Bezeichnungen für ein- und dieselbe chemische Entität.

Method	Example
systematic chemical name	sulfuric acid = hydrogen sulfate = ferrosulfate
common names	aspirin, water
trade names	seroquel = quietapine
company codes	ZD5077 = ICI204636 = ZM204636
abbreviations	DMS for dimethyl sulfate

Tabelle 1: Einige gängige Methoden zur Bezeichnung chemischer Strukturen

Zum Anderen liegen chemische Dokumente in vielen verschiedenen Text- und Bildformaten vor. Insbesondere Patentdokumente sind häufig nur im Bildformat verfügbar, weil dadurch die Integrität und Unveränderlichkeit des Dokuments gewährleistet wird. Einige Patentdokumente liegen zwar als Text vor, dabei handelt es sich jedoch in der Regel um mittels optical character recognition (OCR) konvertierte Bilder. OCR-Dokumente enthalten häufig Übersetzungsfehler, was problematisch ist, da schon kleine Unterschiede in der Schreibweise große Unterschiede in der bezeichneten chemischen Entität bewirken können, wie Abbildung 3 verdeutlichen soll.

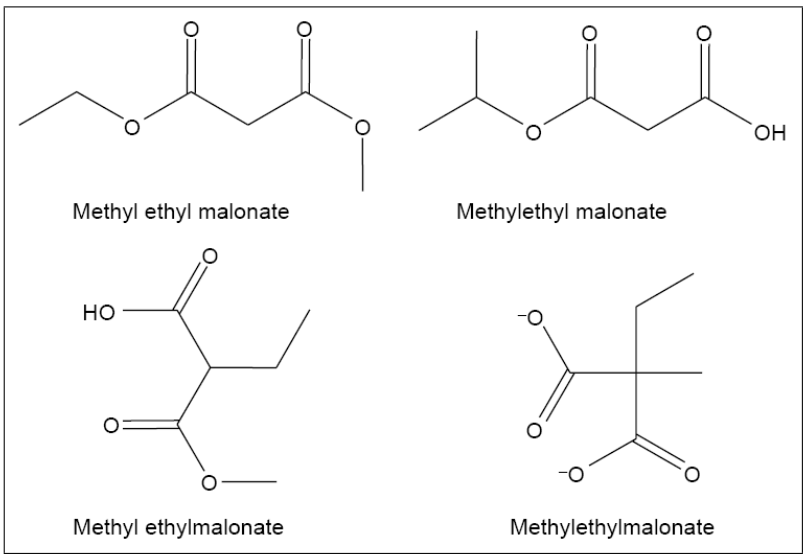


Abbildung 3: Kleine Veränderungen in der Bezeichnung und ihre Auswirkung [Ban06]

Zur Erkennung chemischer Strukturen in Text existieren bereits einige Applikationen. Zu erwähnen sind hier vor allem *ChemAxoms chemicalize.org*[3], *Pipeline Pilot*[4] von *Accelrys* sowie der *ChemExtractor* von *Temis*[5]. Wie in [Ban06] zu lesen ist, arbeiten derzeitige chemische Extraktoren meist regelbasiert.
Im ersten Teil der Studienarbeit werden die oben erwähnten Applikationen zur Extraktion chemischer Begriffe miteinander verglichen. Auf Basis eines noch festzulegenden annotierten Korpus wird sowohl die Performance (Precision, Recall, F-Measure) als auch die Laufzeit der Werkzeuge ermittelt und gegenübergestellt. Anhand dieser Ergebnisse wird eines der Werkzeuge für den Gebrauch ausgewählt.

3.2 Umwandlung in maschinenlesbares Format

Aufgrund der vielfältigen Bezeichnungen für chemische Verbindungen ist eine Konvertierung der mit Hilfe eines externen Tools aus Text extrahierten chemischen Strukturen in ein eindeutiges und maschinenlesbares Format notwendig. Hierfür bietet sich der von der *IUPAC* entwickelte chemische Strukturcode *InChI* (*International Chemical Identifier*) an.
Für die Umwandlung einer Struktur in *InChI*-Schreibweise wird ein von der *IUPAC* unter LGPL veröffentlichter Algorithmus[6] verwendet. Dieses Programm wandelt chemische Strukturinformationen in den drei Schritten Normalisierung, Kanonisierung und Serialisierung in einen eindeutigen *InChI*-Identifikator um. Aufgrund der Eindeutigkeit und Verbreitung eignet sich *InChI* hervorragend für die Integration chemischer Strukturen in einer Datenbank.

3.3 Speicherung und Suche in Datenbank

Nachdem die im Textkorpus lokalisierten chemischen Strukturen in ein standardisiertes Format umgewandelt wurden, werden sie in einer Datenbank gespeichert. Diese Datenbank kann später nach chemischen Strukturen durchsucht werden, was bedeutend schneller ist als eine Durchsuchung des gesamten Textkorpus on-the-fly.
Die Speicherung und Suche extrahierter und in *InChI* konvertierter chemischer Strukturen erfolgt in einer *Oracle* Datenbank mit *Infochems Chemistry Cartridge for Oracle*[7]. Die *Chemisty Cartridge* bietet Funktionalitäten, die für eine Sub- und Superstruktursuche innerhalb der in der Datenbank gespeicherten kodierten Strukturen notwendig sind.
In der Datenbank werden die extrahierten Strukturen in *InChi*-Notation sowie jeglich Links auf Dokumente, in denen die entsprechende Struktur vorkommt, gespeichert. Diese Art der Speicherung als invertierter Index eignet sich hervorragend für eine schnelle Suche von Strukturen in der Datenbank. Ähnlich wie bei *Chemspider* und *PubChem* ist eine Suche innerhalb der Datenbank mit Hilfe einer graphischen Zeichenoberfläche oder einem Eingabefeld für *InChI*-Identifier denkbar.

[3] http://www.chemicalize.org/
[4] http://accelrys.com/products/scitegic/
[5] http://www.temis.com/
[6] Programm verfügbar unter http://www.iupac.org/inchi/
[7] http://infochem.de/en/products/software/iccartridge.shtml

4 Erweiterungen

Eine denkbare Erweiterung für die Suche nach chemischen Strukturen in Text wäre die Einführung eines Ähnlichkeitsmaßes für chemische Verbindungen, die eine Suche nach „ähnlichen“ Strukturen ermöglichen würde. Eine solche Ähnlichkeitssuche wäre sowohl für die Medikamentforschung als auch für Patentanalysen interessant.
Beispielsweise könnten sich Forscher eines Pharmakonzerns fragen „Welche chemische Verbindung haben wir patentiert, die in Struktur oder Wirkung ähnlich ist zu einer wichtigen chemischen Verbindung eines Wettbewerbers?“. Eine weitere mögliche Fragestellung wäre „Welche chemischen Verbindungen sind ähnlich zu einem bekannten Medikament und haben demnach möglicherweise eine ähnliche Wirkung, jedoch ohne die Nebenwirkungen des Medikaments?“

Literatur

[KL04] Ismail Kola, John Landis (2004). *Can the pharmaceutical industry reduce attrition rates?*. In *Nature Reviews Drug Discovery*, 3(8): 711-715.

[Ban06] Debra L. Banville (2006). *Mining chemical structural information from the drug literature*. In *Drug Discovery Today*, 11(1-2): 35-42.

[Sea05] David B. Searls (2005). *Data integration: challenges for drug discovery*. In *Nature Reviews Drug Discovery*, 4(1): 45-58.

[LHWJ06] William Loging, Lee Harland, Bryn Williams-Jones (2007). *High-throughput electronic biology: mining information for drug discovery*. In *Nature Reviews Drug Discovery*, 6(3): 220-230.

[ChEBI08] Kirill Degtyarenko, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickael Guedj, Michael Ashburner (2008). *ChEBI: a database and ontology for chemical entities of biological interest*. In *Nucleic Acids Research*, 36: D344-D350.

[MSim04] Andreas Bender, Robert C. Glen (2004). *Molecular similarity: a key technique in molecular informatics*. In *Organic & Biomolecular Chemistry*, 2: 3204-3218.