



Fachgebiet Wissensmanagement in der Bioinformatik

Prof. Dr. Ulf Leser

Studienarbeit

Verknüpfung funktioneller Annotationen von Genprodukten mit Nachweistexten in wissenschaftlichen Artikeln

Nikolay Damyanliev
damyanli@informatik.hu-berlin.de
Matr.-Nr. 173593

6. Februar 2009

Inhaltsverzeichnis

1	Einleitung	1
2	Hintergrund und bisherige Forschung	1
2.1	BioCreative I - Task 2.1. Ziele und Resultate	2
2.2	Verschiedene Ansätze zur Lösung	3
3	Methoden	5
3.1	Daten und deren Aufbereitung	5
3.2	Erstellung von Proteinlisten	6
3.3	Generierung von GO-Wolken	7
3.4	Bestimmen der relevantesten Textstelle mithilfe eines Sentence-Sliding-Window (SSW)	9
4	Ergebnisse und Diskussion	11
4.1	Nützlichkeit der GO-Wolken	12
4.2	Deutung der Ergebnisstellen	13
4.3	Laufzeit und Speicherverbrauch	16
5	Ausblick und weitere Forschung	16
6	Zusammenfassung	17
	Literatur- und Quellenverzeichnis	18

Abbildungsverzeichnis

Abbildung 1: Beispieleintrag für Proteinannotation aus der PIR-Datenbank (http://pir.georgetown.edu).....	2
Tabelle 2: Die Ergebnisse der verschiedenen Gruppen bei BioCreAtIvE I, Task 2.1. [1]... 3	
Abbildung 3: Beispiel für Text aus einem Artikel vor und nach der Aufbereitung.....	6
Abbildung 4: Beispiel für eine GO-Wolke als XML-Datei.....	8
Abbildung 5: Metriken für Gewichtung.....	9
Tabelle 6: Anzahl von bewerteten Tripeln im Vergleich zu den Ursprungsdaten bei BioCreAtIvE I	11
Tabelle 7: Anzahl von bewerteten Tripeln (unter Benutzung von Okapi-Maß).....	12
Tabelle 8: Anzahl von bewerteten Tripeln in Abhängigkeit von den GO-Wolken (insgesamt 2582 Tripel).....	13
Tabelle 9: Anzahl von bewerteten Tripeln (unter Benutzung von Okapi-Maß).....	14
Abbildung 10: Anzahl der besten fünf Nachweisstellen für alle Tripeln nach ihren Stellen im Text.....	15
Abbildung 11: Anzahl der zu den GO-Termen als relevant bewerteten Nachweisstellen für alle Tripeln nach ihren Stellen im Text	15

1 Einleitung

In dieser Arbeit wird in wissenschaftlichen Artikeln nach Nachweistexten für die funktionellen Annotationen von Genprodukten gesucht. Dazu werden anhand der beim BioCreAtIvE I Wettbewerb, Teilaufgabe 2.1 [1] bereitgestellten Daten mithilfe des Sentence-sliding-window Verfahrens Textausschnitte aus Artikeln bestimmt, in denen die Funktionen von Genprodukten (Proteinen) beschrieben sein sollten. Die Funktionen der Proteine werden in biologischen Datenbanken anhand von Gene-Ontology-Termen (GO-Termen) gespeichert, die eine Beschreibung dieser Funktionen darstellen. Da diese Terme für das in dieser Arbeit besprochene Problem allein keine gute Lösung ermöglichen, wird zu jedem Term eine s.g. Gene-Ontology-Wolke (GO-Wolke) konstruiert, die eine Liste von Wörtern, die nach Relevanz zu dem entsprechenden GO-Term gewichtet sind, darstellt. Unter Verwendung von Listen mit Proteinnamen und diesen GO-Wolken werden danach Textausschnitte aus den Artikeln nach ihrer Relevanz zum Protein und dessen Funktionsbeschreibung (in der Art von GO-Annotationen) bewertet.

Als Eingabedaten werden die Daten, die bei dem BioCreAtIvE I Wettbewerb, Teilaufgabe 2.1 [1] verwendet wurden, benutzt. Ausgabe sind Textausschnitte aus den analysierten Artikeln, die als höchstrelevant bewertet wurden.

2 Hintergrund und bisherige Forschung

Die Molekularbiologie hat in den letzten Jahren eine umfangreiche Menge an Information über Proteine und Gene gesammelt. Die Information, die sich auf die Funktion der Proteine bezieht, wird meistens manuell aus verschiedenen wissenschaftlichen Literaturquellen extrahiert und in biologischen Datenbanken gespeichert, um danach eine wichtige Rolle bei vielen Analyseverfahren in der Bioinformatik zu spielen. Wegen der immer weiter ansteigenden Zahl wissenschaftlicher Artikel, die solche Informationen über Funktionen von Genprodukten enthalten, wächst auch der Bedarf an einem automatisierten Text-Mining-Annotationsverfahren.

Die Gene-Ontology-Terme (GO-Terme) sind das am meisten verbreitete Verfahren zur konzeptionellen Beschreibung von Funktionen von Genprodukten, deswegen werden sie auch meistens als Annotationsvokabular ausgewählt [1]. Die Annotation von Proteinen mit GO-Termen ist eine komplexe Aufgabe, und es wird viel daran gearbeitet und geforscht [6]. Die Annotation wird generell von Kuratoren manuell durchgeführt. Obwohl die manuelle Annotation sehr genau ist [13], ist sie auch sehr zeitraubend, deswegen werden Proteine oftmals auch nicht überwacht mit automatisch generierten Annotationen versehen. Diese sind aber in vielen Fällen mehr oder weniger ungenau, weswegen eine Untersuchung

und Nachprüfung mit Literaturquellen sinnvoll ist. Außerdem wird bei manueller Annotation oft auf ganze Artikel als Nachweisquellen verwiesen (Abb. 1). Es besteht aber die Interesse, die genauen Nachweisstellen in den Artikeln, in denen die Funktion der Proteine extra beschrieben wird, selbst zu kennen. Die Lösung dieses Problems kann bei guten Resultaten auch als Basis für eine bessere automatisierte GO-Annotation genutzt werden.

```
ENTRY      T48678
TITLE      proteasome alpha-1 chain [validated] - Haloferax volcanii
COMPLEX    heterodimer; alpha-1 and beta-1 (PIR:T48677) chain
           [validated; PMID:99412283]
FUNCTION   #description the predominant peptide-hydrolyzing activity of the
           alpha (1)beta(1)-proteasome is cleavage carboxyl to hydrophobic
           residues [validated; PMID:99412283]
```

Abbildung 1: Beispielintrag für Proteinannotation aus der PIR-Datenbank (<http://pir.georgetown.edu>)

Das Hauptproblem dieser Aufgabe besteht im Auffinden des Textfragments, der die GO-Annotation des Proteins beschreibt. Während das Protein selber, bei Vorliegen seines Namens und mit Benutzung von Proteinwörterbüchern mit Synonymen (egal ob fertigen oder selber aus Proteindatenbanken erstellten) in relevanten Artikeln nicht so schwierig nachzuweisen ist, sind die GO-Terme mehr eine konzeptionelle Funktionsbeschreibung des Proteins und nicht für Benutzung in wissenschaftlichen Texten gedacht und stehen sehr oft nur als Beschreibungsterme in Datenbanken, was ihren Nachweis im Text sehr schwierig macht [6].

Die Evaluation der Endresultate bereitet auch gewisse Schwierigkeiten. Da es automatisch nicht entschieden werden kann, welche Textausschnitte relevant sind und welche nicht, müssten alle Resultate manuell durch Kuratoren geprüft werden. In dieser Arbeit besteht keine solche Möglichkeit, deswegen werden hier zur Relevanzprüfung die kurierten Resultate aus dem BioCreAtIvE I Wettbewerb, Task 2.1. benutzt.

2.1 BioCreative I - Task 2.1. Ziele und Resultate

Beim Critical Assessment for Information Extraction in Biology (BioCreAtIvE) Wettbewerb geht es darum, verschiedene Text-Mining-Strategien in Bezug auf biomedizinische Literatur zu evaluieren. Das in dieser Arbeit betrachtete Thema wurde beim BioCreAtIvE I Wettbewerb (2004) als Teilaufgabe 2.1. von ca. 20 Teilnehmergruppen bearbeitet. Anhand gegebener Daten als Tripel (Protein, GO-Term, Artikel) suchten die Teilnehmergruppen für jedes Tripel mit verschiedenen Ansätzen im Artikel nach Nachweistexten, die das Protein und die Beschreibung seiner Funktionen (GO-Term) enthalten und/oder sie in Verbindung bringen. Danach wurden die Resultate

(die zurückgegebenen Nachweistexte) durch Kuratoren vom GO-Annotation-Team (GOA-Team) des Europäischen Instituts für Bioinformatik (EBI) nach Relevanz manuell geprüft und ausgewertet (Tab. 2). Das Ziel war, herauszufinden, ob und welche Text-Mining-Methoden für diese Aufgabe brauchbar sind.

Teilnehmer	Insgesamt ausgegebene Resultate	Relevante Resultate (Precision)	Recall (berechnet auf 1076 Ann.)
Ehrler et al.	1048	268 (25.57%)	24.91%
Krymolowski et al.	1053	166 (15.76%)	15.43%
Vespoor et al.	1057	272 (25.73%)	25.28%
Chiang et al.	251	125 (49.80%)	11.62%
Chiang et al. (2)	45	36 (80.00%)	3.35%
Krallinger et al.	1050	303 (28.86%)	28.16%
Rice et al.	524	59 (11.26%)	5.48%
Ray et al.	413	83 (20.10%)	7.71%
Couto et al.	1050	239 (22.76%)	22.21%

Tabelle 2: Die Ergebnisse der verschiedenen Gruppen bei BioCreAtIvE I, Task 2.1. [1]

Obwohl in den von den Teilnehmern eingereichten Ergebnissen die verschiedensten Text-Mining-Methoden angewendet wurden, waren die Resultate des Wettbewerbs noch ziemlich entfernt davon, die Voraussetzungen für die Anwendung in der Praxis zu erfüllen – eine hohe Precision (über 30%) konnte nur auf Kosten eines sehr niedrigen Recall erreicht werden und das Recall bei allen Teilnehmern lag unter 30% (Tab. 2). Zu den größeren Problemen gehörten wie erwartet die komplexe Natur der Proteinnamen und GO-Terme, sowie der Mangel einer Standardtrainingsmenge.

2.2 Verschiedene Ansätze zur Lösung

Text Mining bietet viele verschiedene Verfahren für die Annotationsaufgabe. Da es eigentlich um Klassifikation von Textausschnitten geht (relevant / nicht relevant), sind auch Klassifikationsverfahren angebracht. Bei BioCreAtIvE I wurden einige davon angewendet. Ray et al. [2] konstruieren für jeden GO-Term eine Menge von „Informationstermen“, wobei sie die Abstracts von Artikeln benutzen, die in verschiedenen Datenbanken mit GO-Annotationen als relevant für den jeweiligen GO-Term angegeben wurden. Diese Informationsterme werden dann verwendet, um die Abschnitte mit der höchsten Relevanz zu bestimmten GO-Termen in den Artikeln zu finden, die in BioCreAtIvE als Ausgangsdaten vorgegeben waren. Der Vorteil dieser Idee besteht darin, dass auf diese Weise mehr Textausdrücke mit Informationsinhalt über den GO-Term gesammelt werden, was zu einer leichteren Suche nach GO-relevanten Absätzen in den Artikeln führt. Einen weiteren Ansatz benutzen Rice et al. [3]: Sie trainieren auf den als Tripel (Protein, GO-Term, Artikel) bereitgestellten Trainingsdaten für jedes Protein und

jeden GO-Term eine Support Vector Machine (SVM), die danach die verschiedenen Textabschnitte bewerten soll. Der Vorteil von diesem Ansatz ist die Ausnutzung der Gen-Ontologien – es wird vermutet, dass nahliegende in der Ontologie Gene ähnliche Funktionen haben. Vespoor et al. [4] konstruieren dagegen mithilfe der bereitgestellten Trainingsdaten zu jedem GO-Term Wortmengen, die dann nach TFIDF mit einer Gewichtung versehen werden. Danach gehen sie die Paragraphen in den Testartikeln durch und bilden den Schnitt zwischen den Wörtern im Paragraphen und den konstruierten Wortmengen. Dieser Schnitt gilt danach als Relevanzfaktor zwischen dem Paragraphen und dem GO-Term. Krallinger et al. [5] betrachten nicht die einzelnen Paragraphen als Nachweistexte, sondern immer 3-4 aufeinander folgende Sätze im Text. Dies ermöglicht es, als Resultat konkretere Nachweisstellen zu bekommen – mit der Idee, dass die Funktion eines Genproduktes meistens in 3-4 zusammenhängenden Sätzen beschrieben wird. So werden genauere Stellen und nicht ganze Paragraphen als Resultate ausgegeben.

Außer den bei BioCreAtIvE I benutzten Methoden wurden später auch andere Ansätze vorgeschlagen, der erwähnenswerteste davon von Gaudan et al. [6]. Sie benutzen wieder nur die GO-Terme als Suchketten, wobei die Textausschnitte nach der Wahrscheinlichkeit gemessen werden, mit der der gesuchte GO-Term in denen auftritt. Das spezifische bei diesem Ansatz ist, dass für jede Kombination (GO-Term, Textausschnitt) drei Parameter berechnet werden - Indiz für den Term, Spezifität der Wörter, die im Term auftreten, und ihre relative Nachbarschaft. Das Indiz wird aus der Wahrscheinlichkeit des Auftretens der Wörter aus dem Term in dem Kontext des Textausschnittes berechnet und die Spezifität ist einfach die Wahrscheinlichkeit des Auftretens der Wörter überhaupt. Die relative Nachbarschaft des Terms misst wie nah beieinander in dem Textausschnitt die Wörter aus dem GO-Term stehen. Im Endergebnis erzielt die Methode eine bessere Precision/Recall-Rate (~34% beide) als die bei BioCreAtIvE I benutzten Ansätze. Das interessante bei Gaudan et al. ist, dass die relativ guten Resultate ohne hinzugefügte Informationen (wie Abstracts relevanter Artikel, Betrachtung von GO-Ontologien o. ä.) erhalten werden. Im allgemeinen spricht das dafür, dass nicht nur die Ansätze, die reines statistikbasiertes Text Mining benutzen (wie der in dieser Arbeit benutzte), bei der gestellten Aufgabe Erfolg haben könnten, sondern dass auch Methoden der Art „pattern matching“ eine angemessene Lösung vorschlagen könnten.

3 Methoden

Das Problem, das in dieser Arbeit betrachtet wird, kann folgendermaßen zusammengefasst werden:

Ist ein Protein P, eine GO-Annotation G und ein relevanter wissenschaftlicher Artikel A gegeben, soll der Nachweistext in A gefunden werden, der P mit G in Verbindung bringt.

Die Suche nach dem besten Nachweistext erfolgt also in zwei Richtungen – es wird zunächst nach Textstellen, in denen das Protein erwähnt wird und danach nach solchen, in denen auch seine Funktion (GO-Annotation) beschrieben wird. Im Schnitt dieser Mengen von Textstellen werden die Textstellen mit der Erwähnung der Beziehung Protein – GO-Term vermutet. Es wird zu jedem Protein eine Liste von Namen und Synonymen, und zu jedem GO-Term eine Menge von gewichteten relevanten Wörtern erstellt. Beide werden danach zur Bewertung der Relevanz der verschiedenen Textausschnitte in dem Artikel benutzt. Als Textausschnitte werden aufeinander folgende Sätze im Artikel betrachtet, wobei die Anzahl der Sätze, die einen Textausschnitt bilden, variabel sein kann. Für die Zwecke dieser Arbeit wurde diese Anzahl auf drei gesetzt.

Das in dieser Arbeit benutzte Verfahren erweitert die Idee von Ray et al. [2] für die Benutzung relevanter Abstracts von Artikeln zur Erweiterung der informationstragenden für die GO-Terme Mengen von Wörtern. Die ausgegebenen Textausschnitte sind wie bei Krallinger et al. [5] keine Paragraphen, sondern aufeinander folgende Sätze im Text.

3.1 Daten und deren Aufbereitung

Als Eingangsdaten wurden hier ausschließlich die bei BioCreAtIvE I – Task 2.1. bereitgestellten Daten benutzt. Diese bestehen aus Tripeln (Protein-ID, GO-Term-ID, Artikel-PMID), Artikeln, GO-Ontologien, Listen von Proteinnamen und Synonymen und Dateien, die zu bestimmten Proteinen und GO-Termen PubMed-IDs relevanter Artikel enthalten. Die Protein-ID entspricht einem bestimmten Eintrag in der UniProt Proteindatenbank, unter dem die Information über dieses Protein gesammelt ist [7]. Durch das sequenzielle Durchsuchen dieser Datenbank wird für jedes Protein eine Liste von Namen und Synonymnamen erstellt, die jeweils in einer separaten Textdatei gespeichert werden. Die Artikel-PMID entspricht einem bestimmten Eintrag in der PubMed-Datenbank mit wissenschaftlichen Artikeln [8]. Die Volltexte der Artikel selbst wurden durch BioCreAtIvE im HTML-Format bereitgestellt. Damit nur die relevante Information in den Artikeln verbleibt und keine Metadaten oder Formatierung die Textauswertung stören, wurden aus jedem Artikel die HTML- und alle weitere Tags, sowie alle Leerzeichensequenzen bis auf einzelne Intervalle entfernt (Abb.3). Da als Textausschnitte immer drei aufeinander folgende Sätze betrachtet werden, wurden auch die einzelnen

Absätze aufgelöst. Beschriftungstexte von Abbildungen wurden als einzelne Sätze im Text gelassen.

Vor der Aufbereitung	<P>The complete primary structure of the human 1(XV) chain consists of 1388 residues, with the following domains: a 25-residueputative signal peptide, a 530-residue N-terminal noncollagenousdomain, a 577-residue collagenous sequence, and a 256-residueC-terminal noncollagenous domain (8).
Nach der Aufbereitung	The complete primary structure of the human 1(XV) chain consists of 1388 residues, with the following domains: a 25-residue putative signal peptide, a 530-residue N-terminal noncollagenous domain, a 577-residue collagenous sequence, and a 256-residue C-terminal noncollagenous domain (8).

Abbildung 3: Beispiel für Text aus einem Artikel vor und nach der Aufbereitung

Die GO-Term-ID entspricht einem bestimmten Eintrag in der Gene-Ontology-Datenbank, in der die Information über alle GO-Terme gesammelt wird [17]. Zu jedem hier benutzten GO-Term wurden aus dieser Datenbank dessen Beschreibungen, Definition, die PubMed-IDs der zu ihm relevanten Artikel und die GO-Term-IDs seiner direkten Oberbegriffe in dem Ontologienetzwerk extrahiert. Die PubMed-IDs der zu einem GO-Term relevanten Artikel werden aus einer bei BioCreAtIvE bereitgestellten Datei gewonnen, in der Bezüge zwischen Proteinen, GO-Termen und Artikeln angegeben werden.

Die bereitgestellten Daten enthalten 2582 Tripel (Protein, GO-Term, Artikel), kombiniert aus 138 verschiedenen Proteinen, 113 Artikeln und 1500 GO-Termen. Jeder Term enthält im Schnitt ca. 60 Wörter. In der bei BioCreAtIvE I, Task 2.1. bereitgestellten Datei werden zu jedem GO-Term im Schnitt 17 Artikel (in der Form von PubMed-IDs) referenziert.

Da die automatische Evaluation der Ergebnisse des hier entwickelten Verfahrens nur auf durch Kuratoren bewerteten Daten möglich ist, werden dafür nur die durch Kuratoren manuell geprüften und evaluierten Resultate der Teilnehmer an dem BioCreative I Task 2.1. genommen und mit den Ergebnissen dieser Arbeit verglichen. Diese sind 1076 Tripel, wobei für fast alle Tripel mehrere bewertete Vorschläge für Nachweisstellen vorhanden sind, insgesamt sind es ca. 10000 bewertete Textausschnitte.

3.2 Erstellung von Proteinlisten

Die Aufbereitung von Listen mit Proteinnamen läuft ziemlich unkompliziert ab. Die UniProt Datenbank wird sequenziell nach den benötigten Proteinen durchsucht

(Durchsuchung nach Protein-IDs) und für jedes Protein werden seine Namen und Synonymnamen extrahiert und in einer Textdatei gespeichert. Die ganzen Namen und Synonymnamen werden als separate Einheiten betrachtet (d.h. bei dem in dieser Arbeit entwickelten Verfahren wird nach Proteinnamen nicht wortweise, sondern phrasenweise gesucht). Die Eingabedaten enthalten 138 verschiedene Proteine, im Schnitt hat jedes Protein 4 Namen und Synonymnamen.

3.3 Generierung von GO-Wolken

Da die Suche nach relevanten zu GO-Termen Textstellen im Text die schwierigste Teilaufgabe ist (die GO-Relevanz der als Resultate in BioCreAtIvE I Task 2.1. ausgegebenen Artikelausschnitte ist zweimal niedriger als die Protein-Relevanz der gleichen – Tab. 6), ist die Verbesserung der Trefferquote für GO-Relevanz von Textausschnitten die beste Methode, bessere Ergebnisse für die gestellte Aufgabe zu erreichen. GO-Terme, also Terme, die die Funktion der Proteine beschreiben, kommen meistens frei im Text oder auch in Zusammenhang mit anderen Wörtern vor. Deswegen wird auch viel darüber geforscht, wie diese Terme in wissenschaftlichen Artikeln besser zu erkennen sind.

In dieser Arbeit wird für jeden GO-Term eine GO-Wolke gebildet, die eine Sammlung von zu einem GO-Term relevanten Wörtern darstellt, wobei diese auch nach Relevanz zum Term gewichtet werden. Die Generierung von GO-Wolken erfolgt in drei Phasen:

- In der ersten Phase wird in die GO-Wolke für einen bestimmten GO-Term die Information über diesen GO-Term aus der GO-Datenbank aufgenommen – nämlich die Beschreibungen, die Definition, die PubMed-IDs der relevanten Artikel, und die GO-Term-IDs der GO-Terme, die diese Instanz umfassen und verallgemeinern, d.h. direkte Oberbegriffe dieser Instanz.
- In der zweiten Phase werden die GO-Term-IDs der Oberbegriffe zusammen mit deren Beschreibungen, Definitionen und PubMedIDs der Artikel, die als relevant angegeben werden, in die GO-Wolke aufgenommen.
- In der dritten Phase werden alle PubMedIDs in der Datei durch die Abstracts der entsprechenden Artikel ersetzt. Alle Abstracts werden aus der PubMed Datenbank für biomedizinische Artikel extrahiert.

Die Wolken werden als XML-Dateien gespeichert, die die ganze Information über den entsprechenden GO-Term enthalten (Abb. 4).

```
<goterm>

<goCode>0006032</goCode>
<goname level="0">chitin catabolic process</goname>
<godef>The chemical reactions and pathways resulting in the breakdown of chitin,
<goname type="1">beta-1,4-linked N-acetylglucosamine catabolic process</goname>
<goname type="1">beta-1,4-linked N-acetylglucosamine catabolism</goname>
<goname type="1">chitin catabolism</goname>
<goname type="1">chitin degradation</goname>
<pmidref>11085997</pmidref>
<pmidref>9748235</pmidref>
<goparent level="1">0006030</goparent>
<goparentname type="0" level="1">chitin metabolic process</goparentname>
<goparentdef level="1">The chemical reactions and pathways involving chitin, a li
<goparentname type="1" level="1">beta-1,4-linked N-acetylglucosamine metabolic pr
<goparentname type="1" level="1">beta-1,4-linked N-acetylglucosamine metabolism</
```

Abbildung 4: Beispiel für eine GO-Wolke als XML-Datei

Nachdem alle Informationen über einen GO-Term (in der Form von einer Sammlung von GO-Termen, Artikel-Abstracts, GO-Definitionen u.a.) auf diese Weise in einer XML-Datei gesammelt wurden, müssen diese Informationen nach Relevanz gewichtet werden. Da meistens die GO-Terme im Artikeltext nicht als ganze zusammenhängende Ausdrücke erscheinen, wird jedes einzelne Wort als separater Informationsträger für den GO-Term betrachtet. Deswegen wird aus der XML-Datei eine Textdatei erstellt, die eine Liste der einzelnen Wörter mit deren Auftretenshäufigkeit in der GO-Wolke darstellt. Die Auftretenshäufigkeit der Wörter aus den relevantesten Ausdrücken (Beschreibung und Definition des GO-Terms) wird dupliziert, damit diese Wörter auf diese Weise bei der Gewichtung ein größeres Gewicht bekommen. Zur Messung des Nutzens der Abstracts der in GO angegebenen relevanten Artikel, sowie der verallgemeinernden GO-Terme werden für die GO-Terme drei Mengen mit GO-Wolken (als Textdateien mit Wörtern) erstellt:

- in der ersten Menge von GO-Wolken werden für die Erstellung der Textdateien alle Abstracts und verallgemeinernden GO-Terme übersprungen
- in der zweiten werden auch alle Informationen nur bis zur ersten Ebene der Verallgemeinerung berücksichtigt
- in der dritten werden alle Informationen bis zur zweiten Ebene der Verallgemeinerung mitgenommen.

Alle Wörter in den Textdateien werden dann nach zwei Metriken gewichtet – nach der Standart-TFIDF-Metrik und nach der Okapi-TFIDF-Metrik (Abb. 5), um zu untersuchen, ob die Wahl der Metrik Einfluss auf die Resultate hat. Bei der Gewichtung wird eine Liste von „Stopwords“ benutzt [9] (die ~200 am häufigsten benutzten Wörter im Englischen werden ausgeschlossen). D.h. aus der Menge der GO-Terme aus den Eingangsdaten werden insgesamt sechs Mengen von Textdateien mit Listen von gewichteten Wörtern erstellt.

$$\begin{aligned}
 \text{a)} \quad w_{i,j} &= tf_{i,j} * \log\left(\frac{N}{df_i}\right) \\
 \text{b)} \quad w_{i,j} &= \log\left(\frac{N - df_i + 0,5}{df_i + 0,5}\right) * \frac{tf_{i,j}}{k((1-b) + b \frac{dl}{adl}) + tf_{i,j}}
 \end{aligned}$$

Abbildung 5: Metriken für Gewichtung.

$tf_{i,j}$ – Anzahl des Wortes i im Dokument j ; df_i – Anzahl der Dokumente, in denen das Wort i vorkommt;

N – Anzahl aller Dokumente; dl – Dokumentlänge; adl – Dokumentlänge im Mittel;

$w_{i,j}$ – Gewicht des Worts i aus dem Dokument j ; k, b - Parameter

(für die Zwecke dieser Arbeit wurden die Parameter $b=0,75$ und $k=2$ gesetzt)

a) Standart-TFIDF-Metrik [10]; b) Okapi-TFIDF-Metrik [11]

3.4 Bestimmen der relevantesten Textstelle mithilfe eines Sentence-Sliding-Window (SSW)

Nachdem die Artikel formatiert, die Listen mit Proteinennamen erstellt und die GO-Wolken mit gewichteten Wörtern generiert wurden, wird für jedes Tripel (Protein, GO-Term, Artikel) auch die Suche nach den Nachweisstellen für die Verbindung Protein – GO-Term in den Artikeln durchgeführt. Mit Hilfe des in Java eingebauten Moduls `java.text.BreakIterator` [12] (wird zur automatischen Erkennung von Sätzen, Wörtern u.a. im Text verwendet) wird der Artikel satzweise durchgesucht. Es werden sukzessiv immer drei aufeinander folgende Sätze (d.h. Sentence-sliding-window – SSW von Größe 3) als Textabschnitt genommen und nach Relevanz ausgewertet, wobei die Auswertung in zwei Phasen stattfindet.

In der ersten Phase wird in jedem der drei Sätze separat nach einem der Namen des Proteins gesucht und zwar nicht nach einzelnen Wörtern, sondern nach dem ganzen Proteinennamen: Bei exakter Übereinstimmung erhält der Satz den höchsten Wert, sonst wird der Abstand zwischen den gefundenen Wörtern berechnet und je nach Größe des Abstandes erhält der Satz einen höheren oder niedrigeren Wert. Die Bewertungen der drei Sätze werden dann aufaddiert – somit wird die Bewertung für die Proteinrelevanz des ganzen Ausschnitts erhalten. Z.B. die Liste mit Namen für das Protein mit der Protein-ID O15085 hat zwei Einträge: „rho guanine nucleotide exchange factor 11“ und „pdz-rhogef“. Jeder Satz im SSW wird separat nach jedem dieser Namen durchsucht. Falls ein Satz einen von den beiden Namen als Wortsequenz enthält, bekommt dieser Satz den höchsten Wert. Falls der Satz die Phrase „rho guanine nucleotide exchange factor 11“ nicht enthält, wird die Phrase in einzelne Wörter separiert („rho“, „guanine“ usw.) und der Abstand zwischen den Wörtern aus der Phrase, die im Satz zu finden sind, gemessen. Je größer der Abstand,

desto niedriger wird der Wert des Satzes. Bei Wortvergleichen wird die Groß- und Kleinschreibung ignoriert.

In der zweiten Phase wird der ganze Ausschnitt nach Relevanz zum GO-Term ausgewertet. Da die GO-Wolke eine Liste von Wörtern ist und keine mehrwörtige Phrasen enthält, werden alle Wörter in dem Ausschnitt nacheinander durchgegangen und deren Gewichtung aus der Wortliste in der Textdatei des GO-Terms aufsummiert (wobei die Groß- und Kleinschreibung ignoriert wird). So wird der GO-Relevanzwert berechnet.

Über die beiden Werte (für Proteinrelevanz und GO-Relevanz) wird eine gewichtete Summe gebildet (hier bekommt die GO-Relevanz eine zweimal größere Gewichtung als die Proteinrelevanz). Das ist auch die Bewertung für den Textausschnitt. Für weitere Untersuchungen wird nicht nur der beste Textausschnitt genommen, sondern die zehn besten, wobei bei jedem Ausschnitt auch vermerkt wird, aus welchem Teil des Artikels er stammt (Anfang, Mitte, Ende).

4 Ergebnisse und Diskussion

Da die automatische Evaluation der Ergebnisse nur auf Daten, die durch Kuratoren bewertet wurden, durchgeführt werden kann, werden auch nur Datensätze betrachtet, die bei BioCreAtIvE durch Kuratoren ausgewertet wurden. Da sich die Formatierung des Textes, sowie die Auswahl von Textausschnitten bei dieser Arbeit von denen der Teilnehmer in BioCreAtIvE unterscheiden, wird ein Maß für den Vergleich der gelieferten Textausschnitte benötigt. Dieses misst die Übereinstimmung der Wörter (ohne Berücksichtigung der Groß- und Kleinschreibung) in beiden Textausschnitten, wobei die Schranke von 70% übereinstimmende Wörter als Grenze für die Gleichheit der Textausschnitte gesetzt wird. Bei Übereinstimmung wird die Bewertung für den in dieser Arbeit gelieferten Textausschnitt übernommen. Da es mehrere Teilnehmer bei dem Wettbewerb gab, die für die gleichen Tripelinstanzen Resultate geliefert haben, wird durch alle bewerteten Resultate für das jeweilige Tripel iteriert und bei Übereinstimmung des Textausschnittes mit mehreren von diesen, wird die höchste Bewertung übernommen. Danach wird die Anzahl von hoch-, mittel- und wenig relevanten Textstellen berechnet und Precision und Recall ausgewertet (Tab. 6).

Relevanz und Precision / Recall	Von Kuratoren bewertete Ursprungsdaten		Hier gefundene und nach den Ursprungsdaten bewertete Textstellen (1. beste Stelle) (unter Benutzung von Okapi-Maß)	
	Zu Proteinen	Zu GO-Termen	Zu Proteinen	Zu GO-Termen
Tripel	2582	2582	2582	2582
Insgesamt vorgeschlagene Textstellen	ca. 10000 - 2104 distinkte und bewertete	ca. 10000 - 2104 distinkte und bewertete	2582	2582
Hoch	1738	950	65	63
Mittel	170	165	21	11
Niedrig	196	989	13	25
Precision	–	–	0,87	0,75
Recall	–	–	0,05	0,07

Tabelle 6: Anzahl von bewerteten Tripeln im Vergleich zu den Ursprungsdaten bei BioCreAtIvE I
 Das Recall wird nur auf den als hoch und mittel relevant bewerteten Daten berechnet (1908 für Proteine und 1115 für GO-Terme).
 Precision wird wie üblich auf den Ergebnisdaten berechnet.
 z.B.: Precision für GO-Terme = $(63 + 11) / (63 + 11 + 25)$
 Recall für GO-Terme = $(63 + 11) / (950 + 165)$

Leider ist diese Art zum Messen der Güte der Ergebnisse dieser Arbeit nicht so aussagekräftig, wie es zu wünschen ist, da sich nicht alle hier zurückgelieferten Textstellen auf diese Weise nach Relevanz zu den Proteinen und GO-Termen messen lassen. Dies ist aber die einzige Weise, auf die eine Abschätzung der Güte des Verfahrens mit diesen Daten möglich ist. Es lässt sich also auch vermuten, dass die hier dargestellte Abschätzung eine Unterschätzung der Güte der entwickelten Methode ist.

4.1 Nützlichkeit der GO-Wolken

Um zu messen, ob große GO-Wolken nützlich sind, wurden drei Arten von GO-Wolken für jeden GO-Term erstellt:

- GO-Wolken ohne Berücksichtigung von verallgemeinernden GO-Termen und von zu den GO-Termen relevanten Abstracts (Typ 1 GO-Wolken),
- GO-Wolken mit Berücksichtigung von GO-Termen der ersten Ebene der Verallgemeinerung und Abstracts, die zu denen relevant sind (Typ 2 GO-Wolken),
- GO-Wolken mit Berücksichtigung von GO-Termen der ersten und zweiten Ebene der Verallgemeinerung und Abstracts, die zu denen relevant sind (Typ 3 GO-Wolken).

Relevanz zu GO-Termen und Precision / Recall	Unter Benutzung von GO-Wolken vom Typ 1 und nur 1. besten Stelle	Unter Benutzung von GO-Wolken vom Typ 2 und nur 1. besten Stelle	Unter Benutzung von GO-Wolken vom Typ 3 und nur 1. besten Stelle
Hoch	23	63	72
Mittel	6	11	19
Niedrig	18	25	34
Precision	0,62	0,75	0,73
Recall	0,03	0,07	0,08

Tabelle 7: Anzahl von bewerteten Tripeln (unter Benutzung von Okapi-Maß) in Abhängigkeit von den GO-Wolken (insgesamt 2582 Tripel)

Das Recall wird nur auf den als hoch und mittel relevant bewerteten Daten berechnet (1115 für GO-Terme).

Precision wird wie üblich auf den Ergebnisdaten berechnet.

z.B.: Precision für Typ 1 = $(23 + 6) / (23 + 6 + 18)$

Recall für Typ 1 = $(23 + 6) / 1115$

Wie die Resultate zeigen (Tab. 7), liefert die Einbeziehung der direkten Oberbegriffe der GO-Terme und der relevanten Abstracts spürbar bessere Resultate (Precision steigt um

mehr als 10%, Recall um 5%); die weitere Verallgemeinerung jedoch zeigt keine große Verbesserung. Es lässt sich vermuten, dass für die allgemeinen GO-Terme oft die gleichen Artikel wie die bei deren Unterbegriffen als relevant angegeben werden – d.h. durch die weitere Verallgemeinerung wird nicht viel neue Information gewonnen.

Eine andere Beobachtung war, dass die Wahl der TFIDF-Metrik für die Berechnung der Wortgewichtung in den GO-Wolken fast keinen Einfluss auf die Resultate hatte (Tab. 8). Sowohl mit der Standard-TFIDF-Metrik, als auch mit der Okapi-TFIDF-Metrik wurden für die meisten Artikel die gleichen Textausschnitte als Resultate geliefert. Dieses ist sogar besser erkennbar, wenn man die besten zehn Ergebnisse für jeden GO-Term beobachtet – auch wenn nicht immer in der gleichen Reihenfolge, waren die mithilfe der einen Metrik berechneten Ergebnisse auch unter den Ergebnissen vorhanden, die mithilfe der anderen Metrik berechnet wurden.

Relevanz zu GO-Termen und Precision / Recall	Top 10 Stellen mit Okapi-Maß		Top 10 Stellen mit Standard-TFIDF-Maß	
	Proteine	GO-Terme	Proteine	GO-Terme
Hoch	367	313	383	345
Mittel	106	81	119	78
Niedrig	101	180	121	200
Precision	0,81	0,68	0,82	0,69
Recall	0,25	0,35	0,26	0,38

Tabelle 8: Anzahl von bewerteten Tripeln in Abhängigkeit von den GO-Wolken (insgesamt 2582 Tripel)
 Das Recall wird nur auf den als hoch und mittel relevant bewerteten Daten berechnet (1908 für Proteine und 1115 für GO-Terme).
 Precision wird wie üblich auf den Ergebnisdaten berechnet.
 z.B.: Precision für GO-Terme mit Okapi-Maß = $(313 + 81) / (313 + 81 + 180)$
 Recall für GO-Terme mit Okapi-Maß = $(313 + 81) / 1115$

4.2 Deutung der Ergebnisstellen

Wie aus den Ergebnissen hervorgeht (Tab. 9), ist die Führung einer Liste mit den besten zehn Ergebnissen (statt nur eins – das beste – zurückzuliefern) ein guter Ansatz. Unter der Betrachtung nur der drei besten Ergebnisse steigt das Recall gleich um das Doppelte, wobei die Precision die gleiche bleibt. Außerdem kann eine solche Liste von drei bis zehn besten Nachweisstellen mehr Information über die Beschreibung der Funktion des Proteins liefern und auch den Annotationskuratoren eine Auswahl von möglichen guten Nachweisstellen anbieten.

Relevanz zu Proteinen / GO-Termen und Precision / Recall	Unter Benutzung nur der besten Nachweisstelle		Unter Benutzung der besten drei Nachweisstellen		Unter Benutzung der besten fünf Nachweisstelle		Unter Benutzung der besten zehn Nachweisstellen	
	Prot.	GOs	Prot.	GOs	Prot.	GOs	Prot.	GOs
Hoch	65	63	179	163	249	216	367	313
Mittel	21	11	46	31	70	48	106	81
Niedrig	13	25	47	78	65	120	101	180
Precision	0,87	0,75	0,83	0,71	0,83	0,69	0,82	0,69
Recall	0,05	0,07	0,12	0,17	0,17	0,24	0,25	0,35

Tabelle 9: Anzahl von bewerteten Tripeln (unter Benutzung von Okapi-Maß) in Abhängigkeit von der Größe der Liste mit den besten Nachweisstellen (insgesamt 2582 Tripel). Das Recall wird nur auf den als hoch und mittel relevant bewerteten Daten berechnet (1908 für Proteine und 1115 für GO-Terme). Precision wird wie üblich auf den Ergebnisdaten berechnet.
z.B.: Precision für GO-Terme mit 1. besten Textstelle = $(63 + 11) / (63 + 11 + 25)$
Recall für GO-Terme mit 1. besten Textstelle = $(63 + 11) / 1115$

Die andere interessante Beobachtung bei den Ergebnisstellen ist ihre Verteilung im Artikel. Von den hier gelieferten Nachweisstellen liegt ungefähr die Hälfte in den ersten 10% und letzten 15% der Artikel (d.h. ganz am Anfang und ganz am Ende). Bei den als gut bewerteten Ergebnisstellen ist der Unterschied sogar noch deutlicher zu sehen – etwa 65% der als gut bewerteten Ergebnisstellen liegen in den ersten und letzten 10% der Artikel (Abb. 10 und 11). Und das ist auch gut nachvollziehbar: Die ersten 10% der Artikel bestehen aus der Einleitung, die letzten aus der Zusammenfassung und / oder Diskussion – und in diesen beiden Teilen eines Artikels wird die wichtigste Information zusammengefasst (also auch die Information über die Proteine, über die in diesem Artikel geschrieben wird, und deren Funktion).

Eine andere Beobachtung unterstützt auch diese Vermutung: Nimmt man einen bestimmten Artikel und betrachtet die Verteilung der Bewertungen der verschiedenen Textausschnitte dort, sieht man das gleiche Bild – die Textausschnitte mit den höchsten Werten liegen meistens am Anfang und am Ende des Artikels, wobei hohe Werte auch irgendwo in der Mitte des Artikels auftreten, meistens werden diese aber durch die Anfang- und Endstellen unterdrückt. D.h. durch die in dieser Arbeit besprochene Methode zur Bewertung von Textstellen erhalten in den meisten Artikeln die Textstellen am Anfang und am Ende die höchste Relevanzbewertung.

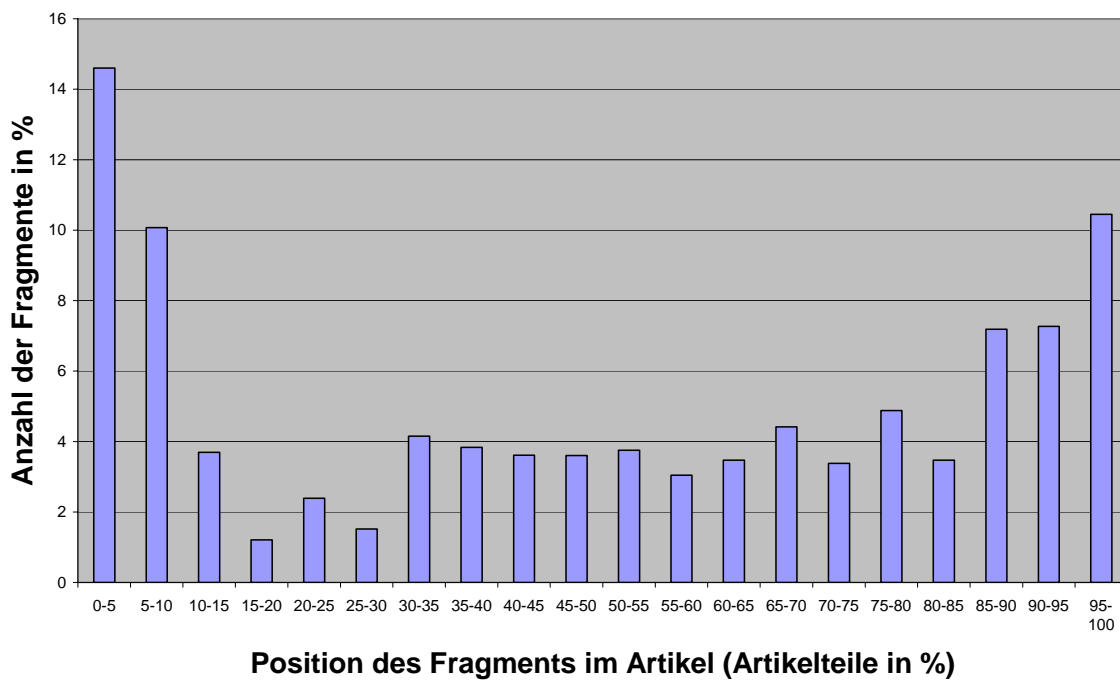


Abbildung 10: Anzahl der besten fünf Nachweisstellen für alle Tripeln nach ihren Stellen im Text

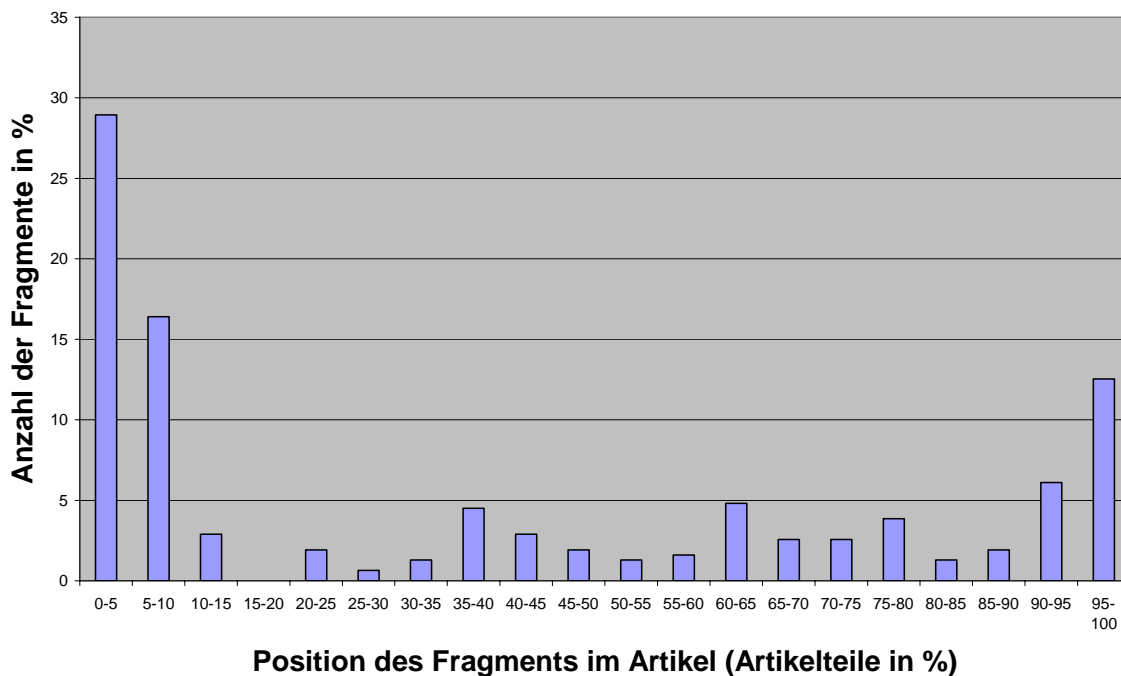


Abbildung 11: Anzahl der zu den GO-Termen als relevant bewerteten Nachweisstellen für alle Tripeln nach ihren Stellen im Text

4.3 Laufzeit und Speicherverbrauch

Erwartungsgemäß können die vielen Daten, die Größe der GO-Wolken und die Größe des Sentence-sliding-window auf die Laufzeit der zur Suche von Nachweisstellen verwendeten Applikationen eine schlechte Auswirkung haben. Es wurden Nachweisstellen für 2582 Tripel (Protein, GO-Term, Artikel) (kombiniert aus 138 verschiedenen Proteinen, 113 Artikeln und 1500 GO-Terme) gesucht. Die zur Evaluation der Ergebnisse benutzten Resultate aus BioCreAtIvE I enthielten ungefähr 10000 von Kuratoren bewertete Nachweistextstellen. Alle Applikationen wurden auf einem AMD Sempron 3000+ (2 GHz) mit 1 GB Arbeitsspeicher getestet.

Die Erstellung von 138 Proteinlisten unter Verwendung der 2 GB großen Textdatei mit Informationen über Proteine, die von der UniProt Datenbank bereitgestellt wird, dauerte ungefähr 3 Minuten. Die Generierung von 1500 GO-Wolken in drei Phasen dauerte insgesamt ungefähr 17 Minuten. Die Erstellung eines Wörterbuchs mit allen Wörtern, die in den 1500 GO-Wolken vorkommen und die Umwandlung der GO-Wolken in Listen mit gewichteten Wörtern dauerten ungefähr 80 Minuten (für GO-Wolken mit Berücksichtigung der Oberbegriffe der ersten Ebene der Verallgemeinerung und der relevanten Abstracts) und waren damit die zeitraubendste Operation. Für größere GO-Wolken stieg die Laufzeit spürbar. Die Auswertung der Ergebnisse dauerte nicht mehr als 20 Minuten, abhängig von der Größe der geführten Liste mit den besten Nachweisstellen. Die Evaluation der Ergebnisse dauerte 1-3 Minuten, wieder abhängig von der Größe der Liste mit den besten Nachweisstellen. Der Arbeitsspeicherbedarf überschritt 600 MB nicht.

Obwohl viele Daten zu verarbeiten sind, halten sich die Ressourcen, die zur Auswertung gebraucht werden, im Rahmen. Dies macht es möglich, zusätzliche Ansätze zu suchen und zu implementieren, die die gelieferten Ergebnisse noch verbessern könnten.

5 Ausblick und weitere Forschung

Die Resultate dieser Arbeit zeigen, dass es schon einige Ansätze zur Verbesserung der Suche von Nachweisstellen für Beschreibungen der Funktionen der Proteine im Text gibt. Im Vergleich zu den Ergebnissen aus BioCreAtIvE I, wird hier mithilfe einer Liste der relevantesten Textstellen ein höheres Recall für die ausgegebenen Nachweisstellen erreicht, wobei die Precision kaum fällt (Tab. 9). Das ist auch zu erwarten - je mehr Vorschläge für eine Nachweisstelle gemacht werden, desto wahrscheinlicher ist es, dass eins der Textausschnitte in der Liste eine gute Nachweisstelle ist. Für die Proteinrelevanz und GO-Term-Relevanz sieht die Verbesserung unterschiedlich aus.

Bei der Suche nach Proteinennamen sind die Resultate, die in dieser Arbeit geliefert werden, nicht viel besser als die bei BioCreAtIvE I erreichten. Das ist aber keine unerwartete

Tatsache - die Proteinrelevanz der meisten Ergebnistextstellen bei BioCreAtIvE I war schon hoch. Außerdem konzentriert sich das hier beschriebene Verfahren auf die Verbesserung der GO-Term-Relevanz der ausgegebenen Nachweisstellen.

Bei der Generierung von GO-Wolken sieht es aber anders aus. Die Einbeziehung von relevanten Abstracts in die GO-Wolken verbessert die Relevanz der Nachweisstellen beträchtlich; desgleichen die Einbeziehung von verallgemeinernden GO-Termen. Leider lässt sich aber durch weitere Verallgemeinerungen nicht sehr viel gewinnen. Ein interessanter Ansatz wäre, Verbalisierungen und Flexione der Wörter in der GO-Wolke miteinzubeziehen. Weitere Möglichkeiten wären nicht nur einzelne Wörter, sondern auch bestimmte Wort-Bigramme oder ganze Wortphrasen (wie Proteinamen oder Namen von biologischen Prozessen) in der Liste zu pflegen, wobei diese ein höheres Gewicht erhalten könnten.

Auch wenn man das aus den GO-Wolken generierte Wörterbuch betrachtet, sieht man, dass auch dort Verbesserungen durchgeführt werden können - die Wort- und Satzerkennungsmodule sind nicht perfekt, d.h. es können bessere NLP-Tools ausprobiert werden, die bei wissenschaftlichen und konkreter bei biomedizinischen Texten eine bessere Satz- und Worterkennung schaffen können.

6 Zusammenfassung

In dieser Arbeit wurde die Ermittlung von Nachweistexten in wissenschaftlichen (konkreter biomedizinischen) Artikeln für funktionelle Annotationen von Genprodukten betrachtet. Diese Aufgabe wurde auch als Teilaufgabe 2.1 bei BioCreAtIvE I für ca. 20 Teilnehmergruppen gestellt, wobei das Ziel war, zu ermitteln, ob und welche verschiedenen Text-Mining-Methoden für diese Aufgabe brauchbar sind. Hier wurde der Sentence-sliding-window Ansatz benutzt, um die verschiedenen Textausschnitte nach Relevanz zu bewerten. Für die Bewertungen wurden Listen von Proteinamen und Listen von nach Relevanz zu GO-Annotationen gewichteten Wörtern (GO-Wolken) benutzt. Aufgrund der Resultate ist zu erkennen, dass die Einbeziehung von Abstracts relevanter Artikel in die GO-Wolken eine gute Auswirkung hat und die Trefferrate steigen lässt. Das Gleiche gilt auch für die Benutzung von Listen der zehn besten Textausschnitte, statt nur einen – den besten – zurückzuliefern.

Literatur- und Quellenverzeichnis

- [1] Blaschke C, Leon EA, Krallinger M, Valencia A: Evaluation of BioCreAtIvE assessment of task 2; BMC Bioinformatics. 2005;6 Suppl 1:S16
- [2] Ray S, Craven M: Learning statistical models for annotating proteins with function information using biomedical text; BMC Bioinformatics. 2005;6 Suppl 1:S18
- [3] Rice SB, Nenadic G, Stapley BJ: Mining protein function from text using term-based support vector machines; BMC Bioinformatics. 2005;6 Suppl 1:S22
- [4] Verspoor K, Cohn J, Joslyn C, Mniszewski S, Rechtsteiner A, Rocha LM, Simas T: Protein annotation as term categorization in the gene ontology using word proximity networks; BMC Bioinformatics. 2005;6 Suppl 1:S20
- [5] Krallinger M, Padron M, Valencia A: A sentence sliding window approach to extract protein annotations from biomedical articles; BMC Bioinformatics. 2005;6 Suppl 1:S19
- [6] S. Gaudan, A. Jimeno Yepes, V. Lee, and D. Rebholz-Schuhmann: Combining Evidence, Specificity, and Proximity towards the Normalization of Gene Ontology Terms in Text; EURASIP Journal on Bioinformatics and Systems Biology; EURASIP Journal on Bioinformatics and Systems Biology; Volume 2008, Article ID 342746, 9 pages
- [7] UniProt Database Documentation - <http://www.uniprot.org/help/>
- [8] PubMed Database Documentation - <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helppubmed.chapter.pubmedhelp>
- [9] University of Glasgow – Department of Computing Science - http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words
- [10] Wikipedia - <http://en.wikipedia.org/wiki/Tf-idf>
- [11] William R. Hersh: Information Retrieval: A Health and Biomedical Perspective; 2003, ISBN 0387955224, pp. 273-274
- [12] Java 2 Platform SE 1.4.2 Documentation - <http://java.sun.com/j2se/1.4.2/docs/api/java/text/BreakIterator.html>
- [13] V. Lee, E. Camon, E. Dimmer, D. Barrell, and R. Apweiler: Who tangos with GOA? Use of Gene Ontology Annotation (GOA) for biological interpretation of ‘-omics’ data and for validation of automatic annotation tools,” Silico Biology, vol. 5, no. 1, pp. 5–8, 2005

- [14] Y. Krymolowski, B. Alex, J. L. Leidner: BioCreative Task 2.1: The Edinburgh-Stanford system; Proceedings of the BioCreative Workshop, Granada.
- [15] F M Couto, M J Silva, P M Coutinho: Finding genomic ontology terms in text using evidence content; BMC Bioinformatics 2005, 6(Suppl 1):S21
- [16] J H Chiang, H C Yu: Extracting Functional Annotations of Proteins Based on Hybrid Text Mining Approaches; Proceedings of the BioCreative Workshop, Granada.
- [17] Gene Ontology Documentation - <http://www.geneontology.org/GO.contents.doc.shtml>