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

Vergleich der gefundenen Entitäten von GeneView und Tees

Exposé zur Studienarbeit

eingereicht von: Sascha Baese
geboren am: 07.04.1986
geboren in: Berlin

Gutachter/innen: Prof. Dr. Ulf Leser

eingereicht am: 14. September 2014

1 Einleitung

Naturwissenschaftliche Forschung produziert eine enorme Menge an Publikationen. Fachartikel stellen den Großteil der Informationsgewinnung im wissenschaftlichen Bereich [5] dar und sind in großen, teilweise frei zugänglichen Online-Datenbanken abrufbar. MEDLINE umfasst beispielsweise mehr als 23 Millionen Zitationen [1] der Bereiche Biochemie, Biomedizin und Biotechnologie und die Anzahl neuer Publikationen wächst exponentiell [6]. Um diese Fülle an zugänglichen Informationen verarbeiten zu können, sind Forscher auf rechnergestützte Prozesse angewiesen [6]. Diese werden unter dem Oberbegriff Text Mining zusammengefasst und basieren prinzipiell auf natural language procession, welches in diesem Bereich als BioNLP bezeichnet wird [1, 4, 6]. Sie beinhalten die automatische Extraktion von Informationen und neuen Erkenntnisgewinn aus Texten durch die Erkennung von Entitäten wie Genen und Proteinen und deren Zusammenhang [1, 6].

Dabei betrachtet man vor allem zwei Klassen von Problemen: Die s. g. named-entity-regognition (NER), bei welcher die Entitäten erkannt werden und die relationship extraction (RE), bei der Zusammenhänge zwischen diesen hergestellt werden [6]. NER erkennt Substrings im Text und kann diese in Kategorien wie Gene oder Proteine einordnen. Dabei können Entitäten auch aus mehr als einem Wort bestehen [6]. Die Zuordnung zu Kategorien kann auch durch Einteilung der vorliegenden Texte in Klassen erfolgen, wodurch beispielsweise Proteinfamilien aufgedeckt werden können [6].

Das zweite Feld im Bereich BioNLP erkennt Aussagen über inhaltliche Zusammenhänge der zuvor aufgefundenen Entitäten [1, 6]. Entitäten können zusammenhängen, wenn sie gemeinsam in einem Satz, Abschnitt, gesamten Text oder auch in einer Textklasse aufttreten [6]. Diese Zusammenhänge werden häufig als Events bezeichnet.

Die folgenden Sektionen werden GeneView und Tees vorstellen, dessen Entitätenerkennungen in dieser Arbeit verglichen werden. Beide Systeme arbeiten auf der Datenbank PubMed, welche frei zugänglich ist und unter anderem MEDLINE zur Literaturgewinnung nutzt.

1.1 GeneView

Durch das web-basierte System GeneView, entwickelt vom Lehrstuhl *Wissensmanagement in der Bioinformatik* des Instituts für Informatik der Humboldt-Universität zu Berlin, können Entitäten durch eindeutige ID's gefunden und Dokumente anhand der Anzahl ihrer Zitationen sortiert werden [3]. GeneView annotiert hierbei alle Datensätze aus PubMed und erkennt Entitäten aus 10 verschiedenen Klassen (u. a. Gene, Medikamente, Krankheiten, Mutationen, Proteine) [3].Insgesamt beihaltet die Datenbank von GeneView ca. 32,8 Millionen Gene, 73,3 Millionen Chemikalien, fast eine Million Einzelnukleotid-Polymorphismen und 3,9 Millionen Protein-Protein-Interaktionen [3]. Die Texte werden über Apache Lucene indiziert und vorab durch eine spezielle Text-Mining-Pipeline weiterverarbeitet, wobei Lucene nicht nur als Speicher, sondern ebenfalls als Such- und Ranking-System agiert [2]. GeneView nutzt mehrere Programme zur NER:

Gene, Einzelnukleotid-Polymorphismen, Gattungen, Chemikalien und Histonmodifikationen werden durch speziell auf den jeweiligen Bereich trainierte Tools erkannt [3]. Ein weiteres Tool erkennt andere named entities wie Zelltypen, Krankheiten, Drogen, Enzyme und Gewebe [3]. Durch diese Vielfalt können gezielte Optimierungseingriffe vorgenommen werden.

1.2 Tees

Das Turku Event Extraction System (kurz: Tees), entwickelt von der *Turku BioNLP Group* des Department of Information Technology der University of Turku, dient speziell dem Auffinden von Events in biologischen Texten [4]. Ihm vorgeschaltet sind verschiedene Systeme zum sentence-splitting und NER [4]. BANNER produziert die Daten zu letzterer [4] und Tees stellt diese zum freien Download zur Verfügung.

2 Ziele

Die vorliegende Arbeit beschäftigt sich mit dem Vergleich der von GeneView und Tees präsentierten Entitäten. Durch statistische Verfahren soll ermittelt werden, ob auftretende Unterschiede signifikant sind. Die genaue Ermmittlung eines aussagekräftigen Tests hierbei Teil der Studienarbeit.

Sollten die Tests ergeben, dass keine signifikanten Unterschiede zwischen der NER von GeneView und Tees liegen, können die Ergebnisse dieser Arbeit als Grundlage für einen Vergleich der RE beider Applikationen genutzt werden.

3 Herangehensweise

Da GeneView und Tees ihre Ergebnisse zum freien Download[1] anbieten, ist der Vergleich der gefundenen Entitäten möglich. Die Daten sind jedoch verschieden repräsentiert und müssen daher zuvor einander angeglichen werden. Dies wird algorithmisch erfolgen und anschließend werden statistische Verfahren zur Auswertung der Daten angewendet.

Die Daten werden mit keinem Goldstandard, sondern auf dem gesamten vorliegenden Datensatz verglichen. Eine zu prüfende Herangehensweise sei wie folgt dargestellt:

Da sowohl GeneView als auch Tees PubMed nutzen, kann ein Test für unverbundene Stichproben erfolgen. Zudem liegt keine Kenntnis über die Verteilung der auszuwertenden Daten vor, weshalb ein nichtparametrischer Test angebracht ist. Da die Einzelvorkommen von Entitäten je Pubmed-Artikel gesucht sind, wählen wir die Differenzen der Treffer von GeneView und Tees als geeignete Parameter. Die Differenzen werden über alle, beiden Systemen bekannte, Pubmed-Artikel berechnet. Die Nullhypothese besagt, dass beide Applikationen gleiche Ergebnisse produzieren und wird mit einem

[1] Tees bietet seine Daten unter `http://evexdb.org/download/` [4] und GeneView unter `http://bc3.informatik.hu-berlin.de/download` (vgl. [3]) zum Download an.

Signifikanzniveau von 5% geprüft. Als Test wird der Wilcoxon-Vorzeichen-Rangtest gewählt, welcher die Differenzen benötigt.

In einem, auf dieses Problem übertragbarem Beispiel, wird der These nachgegangen, ob verschiedene Preisrichter die Teilnehmer an einem Synchronschwimmwettbewerb gleich bewerten. Die Preisrichter arbeiten alle auf dem selben Intervall und betrachten die gleichen Wettkämpfer, weshalb hier die Anwendung des Wilcoxon-Vorzeichen-Rangtests zulässig ist.

Der vorgeschlagene Test ist vor allem interessant, da hier die Richtung und nicht nur die Größe der Differenz betrachtet wird. Demzufolge ist nachvollziehbar, welches der beiden, in dieser Studienarbeit betrachteten, Systeme jeweils eine höhere Anzahl an Treffern gefunden hat.

Literatur

[1] S. Ananiadou, P. Thompson, R. Nawaz, J. McNaught, and D. Kell. Event-based text mining for biology and functional genomics. *Bioinformatics*, 28(16):1–18, 2014. DOI: 10.1093/bfgp/elu015.

[2] P. Thomas, J. Starlinger, and U. Leser. Experiences from developing the domain-specific entity search engine geneview. In *Datenbanksysteme für Business, Technologie und Web (BTW), Magdeburg, Germany*, 2013.

[3] P. Thomas, J. Starlinger, A. Vowinkel, S. Arzt, and U. Leser. Geneview: A comprehensive semantic search engine for pubmed. *Nucleic Acids Research*, 40, Web Server issue:W585–W591, 2012. DOI: 10.1093/nar/gks563.

[4] S. Van Landeghem, J. Björne, C.-H. Wei, K. Hakala, S. Pyysalo, S. Ananiadou, H.-Y. Kao, Z. Lu, T. Salakoski, Y. Van de Peer, and F. Ginter. Large-scale event extraction from literature with multi-level gene normalization. *PLoS ONE*, 8(4):e55814, 2013. DOI: 10.1371/journal.pone.0055814.

[5] J. Wilbur, L. Hirschman, and A. Valencia. Evaluation of text-mining systems for biology: overview of the second biocreative community challenge. *Genome Biology*, 9(Suppl 2):S1.1 – S1.9, 2008. DOI: 10.1186/gb-2008-9-S2-S1.

[6] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. Cohen. Frontiers of biomedical text mining: current progress. *Brief Bioinform*, (8)5:358–375, 2007. DOI: 10.1093/bib/bbm045.