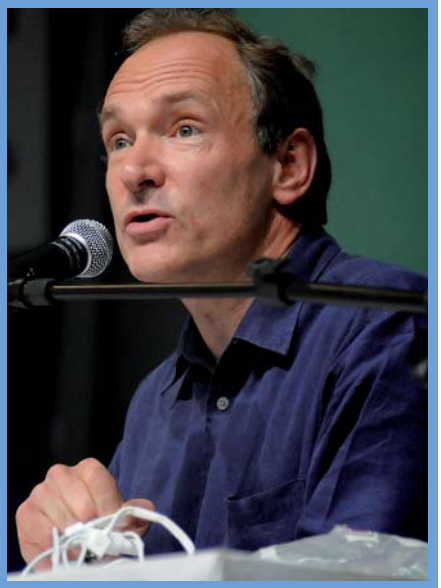
Semantisches Textmining für Ärzte und Kliniken

Digitale Leseratten

Täglich werden tausende neue Artikel publiziert, Krankenakten geschrieben oder E-Mails verschickt. Wie soll man da noch den Durchblick behalten? Dank semantischem Textmining lernen Computer, Texte zu lesen und zu verstehen, und beantworten schnell und vollständig unsere Fragen.

„Wozu die zahllosen Bücher und Bibliotheken, deren Besitzer in seinem ganzen Leben kaum die Titel gelesen hat? Den Lernenden belastet die Masse, anstatt ihn zu belehren, und viel besser ist es, dich wenigen Autoren anzuvertrauen, als durch viele dich zu verirren.“



Sir Tim Berners-Lee (Bildquelle: Wikipedia)

Diese Aussage ist so aktuell, dass man kaum glauben mag, dass sie vor zweitausend Jahren von dem römischen Philosophen Seneca niedergeschrieben wurde. Aber warum verirren wir uns? Wir besitzen doch heutzutage Computer, die selbst das gesamte Internet in Windeseile durchsuchen können. Es ist der Berg an Suchergebnissen, der uns buchstäblich erdrückt.

Ein Beispiel: An einer Klinik wird eine neue, seltene Krankheit untersucht. Die einzelnen Symptome sind nicht ungewöhnlich, aber ihre Kombination lässt aufhorchen: „Schlafstörung verursacht durch häufigen Harndrang“ oder „Angstzustände bei Dunkelheit“. Die geringe Anzahl diagnostizierter Fälle stellt für die Ärzte ein großes Problem dar. Eine systematische Aufarbeitung aller verfügbaren Archive könnte weitere, nicht diagnostizierte Fälle dieser Krankheit aufdecken, das Symptomprofil schärfen und Rückschlüsse auf Ursachen erlauben, aber wer soll das tun?

Bei Pubmed gibt es mehr als zwei Millionen Fachartikel zu Krebs.

Die Suchfunktion unserer Computer kann nur nach einzelnen Stichworten suchen. Begriffe wie „Schlafstörung“, „Harndrang“ oder „Angstzustände“ kommen in vielen, sogar in extrem vielen Dokumenten vor. Das hilft uns nicht weiter. Was im Kontext einer Archivabfrage noch harmlos wirkt, wird eindeutig zu einem ernsthaften Problem, wenn wir mit großen Datenbanken wie beispielsweise der medizinischen Weltliteratur in „Pubmed“ arbeiten. Allein zu Krebs gibt es dort mehr als zwei Millionen Fachartikel.

Auf die Frage „Wodurch wird Krebs verursacht?“ erhalten wir Lesestoff, an dem wir Jahrhunderte sitzen würden. Was wir wirklich brauchen, sind Computer, die nicht nach Stichworten, sondern nach Aussagen suchen, die also den Inhalt eines Satzes „verstehen“. Und die nicht Dokumente zurückliefern, sondern die gestellte Frage beantworten: „durch UV-Strahlen“, „durch Dioxin“, „durch Viren“. Solche Computersysteme gibt es aber nicht. Oder gab es bis vor kurzem nicht.

In den letzten Jahren nahm eine Vielzahl wissenschaftlicher und kommerzieller Forscherteams die Herausforderung begeistert auf – die Wissenschaftszeitschrift Nature spricht gar vom nächsten Google – und entwickelt eine ganze Reihe von Technologien und Standards für die semantische Analyse unstrukturierter Daten. „Semantische Analyse“, „unstrukturierte Daten“? Semantische Analyse heißt, dass die Bedeutung erkannt wird. Und zu den unstrukturierten Daten zählen alle Informationen, die in einer nicht klar definierten Form vorliegen. Dies gilt leider für fast alle Informationen, denen wir täglich begegnen, insbesondere im medizinischen Umfeld. Publikationen, Patientendaten, E-Mails, PDFs, Vorträge, Notizen jeglicher Art, sie alle gehören dazu. Sie sind für Menschen gemacht und daher erst einmal unstrukturiert.

Die Menge unstrukturierter Daten hat in den letzten Jahren massiv zugenommen. Betrachtet man beispielsweise die Anzahl neu erschienener biomedizinischer Publikationen – 2010 waren es mehr als eine Million – kann man sich leicht vorstellen, dass man mit klassischem Durchlesen

und Sich-Merken nicht weit kommt. Die schiere Textmenge ist auch für eine Gruppe von Lesern nicht mehr fassbar und schon gar nicht reproduzierbar. In anderen Worten: Wir produzieren täglich mehr Information, als wir nützen können, und das bei teilweise horrenden Kosten, man denke nur an wissenschaftliche Experimente.

Semantische Technologien sind entwickelt worden, um diese Wissensschätze zu heben und für uns bereitzustellen. Welches Wissen ist in einem Satz enthalten, welche Beziehungen werden postuliert, wovon handeln die Worte? Damit ein Computer solche Interpretationen bewerkstelligen kann, müssen ihm natürlich geeignete Vokabulare zur Verfügung gestellt werden. In sogenannten Ontologien haben Wissenschaftler genau definiert, was bestimmte Worte bedeuten und welche synonymen Schreibweisen es davon gibt. Ein konkretes Beispiel wäre PTEN, das in die Kategorie der Proteine, Untergruppe cytoplasmatische Proteine, gehört und auch *Phosphatase and Tensin Homolog* genannt wird.

Semantische Technologien sollten jedoch nicht nur einzelne Begriffe korrekt einordnen können. Die wahre Stärke dieser Programme zeigt sich erst, wenn auch die Beziehungen zwischen den einzelnen Wörtern richtig interpretiert wird.

Ein Beispiel eines solchen semantischen Textmining Systems wurde von uns im Rahmen der Helmholtz-Allianz Systembiologie entwickelt. Konkret wurde die gesamte biomedizinische Literatur – etwa 20 Millionen wissenschaftliche Artikel – semantisch analysiert. Die einzelnen Schritte der automatischen Textinterpretation sind relativ komplex, lassen sich aber vereinfacht folgendermaßen beschreiben: Das Programm zerlegt die Artikel in einzelne Sätze und weist den Satzteilen semantische Rollen wie zum Beispiel Subjekt, Objekt oder Prädikat (Verb) zu. Danach werden die Satzteile mit Begriffslisten abgeglichen – die Interpretation ist fertig. Konkret könnte das Resultat nun so ausschauen: Das Subjekt ist „PTEN", das Objekt ist „Krebs", das Prädikat gehört zur Gruppe „Inhibition", also „PTEN verhindert Krebs".

Wir produzieren weit mehr Informationen als wir nützen können.

In einer Datenbank, die biomedizinische oder andere Daten als semantische Relationen speichert, können Fragen nun viel gezielter und reproduzierbarer beantwortet werden als mit klassischen Suchmaschinen. Machen wir ein Beispiel: Wir wollen wissen, bei welchen Krankheiten PTEN ein tatsächlicher oder möglicher Biomarker ist. Geben wir nun bei Google oder Pubmed die Stichworte „PTEN Biomarker" oder „PTEN regulates" ein, so erhalten wir eine sehr große Anzahl an Dokumenten, die alle die gesuchten Stichworte enthalten. Es ist nun dem Suchenden überlassen, sich durch all die Treffer durchzuarbeiten. Da wir aber kaum mehr als die ersten 30-40 Treffer überfliegen und die restlichen Tausenden von Dokumenten unbeachtet lassen, ist unser Resultat in jedem Fall stark zufällig und vor allem unvollständig. Gerade diese Unvollständigkeit ist ein großer Nachteil klassischer Suchmaschinen gegenüber semantischen Technologien. Bei Google oder Pubmed müssten wir uns durch hunderte von Dokumenten durcharbeiten, um zu entdecken, dass PTEN nicht nur bei Krebs ein Biomarker ist, sondern auch bei Parkinson eine wichtige Rolle spielt und potenziell als Biomarker genutzt werden könnte.

Bei einer semantischen Suchabfrage hingegen werden die Resultate gruppiert, zusammengefasst und in einer strukturierten Übersicht präsentiert. Auf einen Blick sehen wir alle Krankheiten, in denen PTEN eine Rolle spielt. Nebst Krebs, Diabetes und Asthma erscheint auch Parkinson. Erst wenn wir die einzelne Krankheit anklicken, erscheinen die zu Grunde liegenden Dokumente, die dann eine genauere Beurteilung der Aussage zulassen.

Genau diese Werte – Vollständigkeit, Relevanz, Übersichtlichkeit und auch Geschwindigkeit – sind Gründe, warum viele experimentelle Biologen und Kliniker für komplexere Fragestellungen vermehrt auf semantische Lösungen zurückzugreifen. Sei es, dass Gene einer GWAS-Studie interpretiert und miteinander in Verbindung gebracht werden müssen, sei es dass neue Biomarker gesucht, Wirkungsmechanismen von Medikamenten verstanden oder heikle Patientendaten aufgearbeitet werden müssen. Semantische Text Mining Techniken konnten in den letzten Jahren stetig wachsende Beliebtheit in der Wissenschaft und im Gesundheitswesen verzeichnen. Es gibt viel zu tun und bleibt spannend, auch für uns.

Dr. Volker Stümpflen, Dr. Michael Greeff
Inst. für Bioinformatik und Systembiologie
v.stuempflen@helmholtz-muenchen.de

Zum Testen gibt es im Internet ein kostenloses Programm EXCERBT, das am Helmholtzzentrum München entwickelt wurde (*http://mips.helmholtz-muenchen.de/geknowme/web/excerbt*). Es funktioniert ähnlich einfach wie Google: Nach Eingabe eines Suchbegriffs (z. B. *NF kappa B*) erhält man eine Trefferliste in Form eines strukturierten Baums mit Knoten für Aktivierung, Hemmung usw. Daraus kann man ein „semantisches Netz" konstruieren.