Destillation von Wissen aus Texten und dem Web

Von Gerhard Paaß, Fraunhofer IAIS

Erschienen in dok - Technologien, Strategien & Services für das digitale Dokument, Okt. 2010, S.6.

In den letzten Jahren hat sich das Internet zu einer Informationsquelle ersten Ranges entwickelt. Zum Auffinden von Informationen werden meist einige Stichworte in eine Suchmaschine, z.B. Google, eingegeben. Diese liefert dann eine Liste von Internetseiten zurück, in denen die gewünschten Stichworte auftauchen.

Die Resultate der Suchmaschinen sind unbefriedigend, falls mehrere Begriffe gleichzeitig gesucht werden. Stellt man in Google z.B. die Anfrage „Medikament gegen Bluthochdruck für Diabetiker“, so erhält man Webseiten, in denen die Worte Medikament, Bluthochdruck und Diabetiker gleichzeitig vorkommen. Allerdings werden sinngleiche Wörter wie etwa Arznei, Hypertonie oder Zuckerkrankheit ignoriert. Zudem wird in den meisten gefundenen Seiten darüber berichtet, dass Medikamente gegen Bluthochdruck als Nebenwirkung selbst Zuckerkrankheit verursachen können. Die durch die Worte „gegen“ und „für“ in der Suchanfrage ausgedrückten Zusammenhänge werden nicht berücksichtigt und nur ein kleiner Bruchteil der Treffer schlägt geeignete Medikamente gegen Bluthochdruck für Diabetiker vor.

Als einen ersten Schritt zur Darstellung inhaltlicher Zusammenhänge zwischen Stichworten wurde die Idee eines „Internets der Bedeutungen“ (Semantic Web) entwickelt. Zu jeder Webseite wird die Bedeutung des Textes in einer standardisierten Form mit einheitlichen Begriffen abgespeichert. Zentral ist hierbei die Möglichkeit, inhaltliche Relationen zwischen zwei Konzepten auszudrücken (z.B. „*Betablocker* ist wirksam gegen *Bluthochdruck*“). Es lässt sich zeigen, dass auf diese Weise nahezu beliebige Textinhalte semantisch eindeutig und vollständig festgehalten werden können. Diese semantischen Informationen können dann von Maschinen interpretiert und automatisch maschinell weiterverarbeitet werden.

Da eine manuelle semantische Annotation sehr aufwendig ist, wurden in den letzten Jahren automatische Verfahren entwickelt (z.B. im Theseus-Programm der Bundesregierung), die verschiedene Aspekte der Bedeutung von Texten erfassen und in die standardisierte Darstellung überführen können. Nach dem Ansatz des „Maschinellen Lernens“ analysieren sie „Trainingsmengen“ von Texten, in denen die Begriffe manuell durch ihre Bedeutungen annotiert wurden. Die Verfahren sind in der Lage eine Vielzahl von Mustern in den Texten zu erkennen und mit einer hohen Wahrscheinlichkeit die gesuchten Bedeutungen zu identifizieren. Ihr Genauigkeitsniveau hängt dabei von der Komplexität der Aufgabe sowie der Qualität und dem Umfang der Trainingsdaten ab.

Es hat sich ein allgemeiner Workflow herausgebildet, in dem die Texte mit Annotationen versehen und für semantische Nutzerabfragen aufbereitet werden. Dieser Workflow umfasst vier Schritte:

1. Extraktion von Bedeutungen für einzelne Worte und Begriffe. Hier werden insbesondere unterschiedliche Arten von Eigennamen (Personen, Orte, etc.) sowie Fachbegriffe mit mehreren Worten identifiziert.
2. Disambiguierung, d.h. Auflösung von Mehrdeutigkeiten und Zuordnung zu Konzepten. Einerseits gibt es verschiedene Worte mit der gleichen Bedeutung (Bluthochdruck,

Hypertonie) oder aber auch verschiedene Bedeutungen für das gleiche Wort (Finanzinstitution Bank, Parkbank). Worte mit nahezu gleicher Bedeutung bilden ein Konzept. Die Auflösung von Mehrdeutigkeiten beruht auf dem Vergleich der Nachbarworte.

3. Extraktion von Relationen zwischen zwei Konzepten. Ausgehen von den Worten eines Satzes und der syntaktischen Analyse des Satzbaus können Verfahren der Relationsextraktion nach einem Training mit geeigneten Trainingsdaten Zielrelationen zwischen Konzepten identifizieren.
4. Abspeichern der gefundenen Fakten in einer Datenbank oder Suchmaschine. Die gefundenen Annotationen, Konzepte und Relationen bilden – zusammen mit extern vorgegebenen Informationen – ein Netzwerk (Ontologie) von logischen Relationen (z.B. A ist Oberbegriff von B, „Betablocker ist wirksam gegen Bluthochdruck“). Die gefundenen Fakten kann man zusammen mit den zu Grunde liegenden Texten in einer Suchmaschine indexieren, oder sie in speziellen Ontologie-Datenbanken nach dem RDF- bzw. OWL-Standard ablegen.

Automatische semantische Annotationen sind mittlerweile recht leistungsfähig. Während die Identifikation von Eigennamen mit einer Sicherheit um die 90% möglich ist erreicht man bei der Extraktion von Relationen in Freitexten Genauigkeiten im Bereich von 80-90%. Bei weiterer Verbesserung der Leistungsfähigkeit haben sie ein hohes Potential in der Analyse von Meinungen in sozialen Netzen, der Suche nach Produkten in Publikationen und Patenten oder aber auch der Platzierung von Werbung zu redaktionellen Beiträgen.