Don Giovanni ist kein Rückschlagspiel

Wie Wikipedia helfen kann, Freitexte zu vergleichen

von Ute Schmid und Matthias Düsel

Eine Bank ist nicht nur zum Sitzen da. Vergleicht ein Computer zwei Texte miteinander, können sich semantische Mehrdeutigkeiten einschleichen, die nur durch recht aufwändige Verfahren vermieden werden können. Bei kurzen Texten ist das Kategoriensystem der Wikipedia-Artikel eine Hilfe. Bamberger Informatiker nutzen es in einem Pilotprojekt, um Senioren den Alltag zu erleichtern.

Was haben Tennis und Fußball oder Schach und Schafkopf gemeinsam? Menschen organisieren ihr semantisches Wissen häufig in hierarchischen Kategoriensystemen, die ihnen helfen, Dinge, Ereignisse oder den Inhalt sprachlicher Äußerungen zu klassifizieren und nach Ähnlichkeit zu vergleichen: Tennis und Fußball sind beides Ballsportarten, Schach und Schafkopf sind Spiele.

Möchte man Freitexte per Computer nach ihrer Ähnlichkeit vergleichen, fehlt dieses semantische Wissen erst einmal. Im einfachsten Fall werden Texte nach ihrer lexikalischen Ähnlichkeit verglichen. Dabei wird geprüft, wie viele Schlüsselwörter die Texte gemeinsam haben.

Texte syntaktisch und semantisch vergleichen

Allerdings können bei einem solchen rein syntaktischen Abgleich passende Dokumente übersehen und unpassende Dokumente als ähnlich klassifiziert werden. Der erste Fall tritt ein, wenn mit verschiedenen Begriffen auf denselben Sachverhalt referenziert wird, wenn also Synonyme wie *Apfelsine* und *Orange* verwendet werden. Der zweite Fall tritt ein, wenn das gleiche Wort auf verschiedene Dinge referenziert, wenn also Polyseme wie *Bank* verwendet werden, die einmal ein Geldinstitut und einmal eine Sitzgelegenheit bezeichnen.

Um mit solchen Fällen umzugehen, kann die einfache lexikalische Analyse erweitert werden – etwa indem Information aus einem Synonym-Lexikon mit einbezogen wird. Anderseits wurden verschiedene Ansätze entwickelt, die direkt die Ermittlung der semantischen Ähnlichkeit von Texten im Fokus haben. Im Wesentlichen lassen sich zwei Herangehensweisen unterscheiden: Verfahren, in denen Bedeutung implizit ermittelt wird, und solche, die auf einer expliziten Modellierung semantischen Wissens beruhen. Implizite Verfahren nutzen Methoden der Statistik und des maschinellen Lernens. Damit diese Ansätze gut funktionieren, ist es in den meisten Fällen notwendig, dass viele – und nicht zu kurze – Texte in die Analyse mit einbezogen werden.

Taxonomien und Ontologien

Um Wortbedeutung explizit in den Ähnlichkeitsvergleich miteinzubeziehen, muss Wissen über die Begriffe, die durch die Worte repräsentiert werden, in einem Modell abgelegt werden. Eine einfache, und in der Künstlichen-Intelligenz-Forschung schon seit langem eingeführte Art der Wissensrepräsentation sind Taxonomien, also Ober-/Unterbegriff-Beziehungen.

Wie genau Wissen in Taxonomien abgebildet wird, hängt vom Anwendungskontext ab. Beispielsweise könnte eine Taxonomie für *Sport* eine Kategorie *Ballsport* und eine Kategorie *Leichtathletik* enthalten. Ballsport könnte in *Zwei-Personen-Spiele* und *Mannschaftsspiele* unterteilt werden. Kommt nun in einem Text das Wort *Tennis*, im anderen das Wort *Fußball* vor, wären beide über zwei Hierarchieebenen miteinander verbunden. Fußball und Basketball wären sich ähnlicher als Fußball und Tennis, da beides Mannschaftssportarten sind. Hätte man aber stattdessen die Ballsportarten in *Hallen-* versus *Rasensport* unterteilt, wären sich Basketball und Fußball unähnlicher. Ausdrucksstärker als Taxonomien sind Ontologien, die im Bereich des *semantic web* intensiv erforscht werden. Hier kann Wissen nicht nur hierarchisch, sondern in einem Netzwerk organisiert werden.

Wird explizites semantisches Wissen zum Vergleich von Texten einbezogen, bietet das den Vorteil, dass Worte Begriffen zugeordnet werden und damit nicht im Text genannte Information mit in die Analyse einfließen können.

Semantische Ähnlichkeit via Wikipedia

Hat man nur wenige und kurze Texte zur Verfügung, sodass indirekte Methoden keine brauchbaren Ergebnisse liefern, und möchte man dennoch die explizite Modellierung umgehen, kann Wikipedia sehr nützlich sein. Auf jeder Wikipedia-Seite werden unten nach dem Schlüsselwort Kategorien-Begriffe angegeben, in die das auf der Seite erläuterte Thema eingeordnet werden kann. Beim Lemma *Tennis* findet sich die Kategorie *Rückschlagspiel*, geht man auf die Seite *Rückschlagspiel*, findet der User die Kategorie *Ballsportart*. Diese Methode – Tiefensuche im Kategoriengraph – haben Bamberger Informatiker verwendet, um ein System zu entwickeln, das Seniorinnen und Senioren dabei unterstützen soll, aktiv und mobil zu bleiben. Diese Mobilitätsplattform heißt *MoNA*, weitere Informationen dazu finden sich im Textkasten.

In einer webbasierten Anwendung sollen Senioren mit *MoNA* eine einfache und intuitiv zu bedienende Oberfläche nutzen können, die ähnlich wie eine Pinnwand im Supermarkt funktioniert: Sie können Angebote sowie Anfragen einstellen. Während man aber bei einer Pinnwand alle Aushänge durchliest und dann entscheidet, welche Angebote oder Anfragen zu den eigenen Interessen passen – also die semantische Ähnlichkeit von eigenem Interesse und Aushang ermittelt –, soll das System automatisch die passenden Angebote zu einer Anfrage liefern.

Beispieleinträge sind in der Abbildung gezeigt: Die beiden Anfragen „Ich sehe mir am Sonntag *Don Giovanni* an. Wer begleitet mich?“ und „Ich gehe gerne in die Oper“ hätten bei einem lexikalischen Vergleich die Ähnlichkeit 0. Dagegen würde der unpassende Eintrag Nummer 6 gefunden. Nutzt man nun die Wikipedia-Kategorien, können beide Listen durch die zugeordneten Kategorien erweitert werden. Für „Don Giovanni“ findet sich unter anderem „Oper von Wolfgang Amadeus Mozart“, und damit haben beide Texte nun einen gemeinsamen Begriff.

ID	Eintrag	Autor
0	Ich sehe mir am Sonntag Don Giovanni an.Wer begleitet mich?	Person0
1	Ich brauche jemanden der mir beim Einkaufen hilft.	Person1
2	Wer möchte mit mir in ein Konzert gehen?	Person2
3	Ich gehe gerne in die Oper.	Person3
4	Ich gehe heute Abend in die Oper und suche noch jemanden der mitkommt.	Person4
5	Ich interessiere mich für alle Arten von Musikveranstaltungen.	Person5
6	Ich gehe gerne ins Kino und suche eine Begleitung	Person6
7	Ich suche jemanden, der mir mir shoppen geht.	Person7
8	Ich suche eine neue Lauf-Gruppe.	Person8
9	Wer geht mit mir joggen?	Person9
10	Ich gehe gerne spazieren.	Person10
11	Partner für Nordic Walking gesucht.	Person11

Literaturempfehlung

Ute Schmid u. a.: How similar is what I get to what I want: Matchmaking for Mobility Support. In: *Computational Approaches to Analogical Reasoning: Current Trends.* Heidelberg: Springer 2013, S. 263–287.

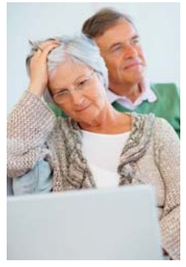
Matthias Düsel: Semantisches Matching von Freizeitaktivitäten mittels Wikipedia-basierter Kategorisierung. Masterarbeit im Studiengang Angewandte Informatik. Fakultät Wirtschaftsinformatik und Angewandte Informatik, Universität Bamberg, Dezember 2013.

Christoph Schlieder, Ute Schmid u. a.: Assistive Technology to Support the Mobility of Senior Citizens. In: KI-Künstliche Intelligenz 27 (2013), H. 3, S. 247–253.

Natürlich funktioniert auch dieser Lösungsansatz nicht in jedem Fall perfekt. Beispielsweise wird Eintrag 5 aus der Abbildung nicht gefunden, wenn man die Suche im Kategoriengraph auf wenige Ebenen begrenzt. Vergleicht man jedoch die Suche im Kategoriengraph mit einfacheren Verfahren, so liefert dieser Ansatz häufig die besten Ergebnisse bezüglich *Recall* – die Wahrscheinlichkeit, mit der ein relevanter Eintrag gefunden wird – und *Precision* – die Wahrscheinlichkeit, mit der ein gefundener Eintrag relevant ist. Und schon kann die gemeinsame sonntägliche Opernfahrt organisiert werden.

Das Projekt EMN-Moves

Das vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Verbund-Projekt *Die Europäische Metropolregion Nürnberg macht mobil durch technische und soziale Innovationen für die Menschen in der Region* lief von November 2011 bis Oktober 2014. Es war Teil der BMBF-Initiative *Mobil bis ins hohe Alter – nahtlose Mobilitätsketten zur Beseitigung, Umgehung und Überwindung von Mobilitätsbarrieren* in der bundesweit zehn Großprojekte gefördert wurden. Für die Universität Bamberg waren Prof. Dr. Christoph Schlieder und Prof. Dr. Ute Schmid sowie die Mitarbeiter Dr. Klaus Stein und Michael Munz beteiligt. Gemeinsam entwickelten sie die Mobilitätsplattform *MoNA*, die Seniorinnen und Senioren in ihrer Mobilität in ihrem Wohnumfeld unterstützt. Zentrales Anliegen war dabei Mobilität als soziale Aufgabe zu betrachten, in die soziale Organisationen, Wohnungsbaugesellschaften sowie Anwohner verschiedener Altersgruppen eingebunden sind.



www.uni-bamberg.de/kogsys/research/projects/bmbf-project-emn-moves-match-making

Don Giovanni is not a Racquet Sport

How Wikipedia can help to compare free texts

A bank is not just the side of a river. When a computer is used to compare two texts, semantic ambiguity can creep in, and avoiding this means implementing highly complex processes. Wikipedia's article categorisation system can be helpful when dealing with short texts, and computer scientists at the University of Bamberg are using it in a pilot project aimed at making everyday life easier for the elderly.