Wikipedia: Wissen für die KÜNSTLICHE INTELLIGENZ

Wenn ein Computer sich Allgemeinwissen aneignen soll, ist die Online-Enzyklopädie Wikipedia eine überaus reichhaltige Quelle – vorausgesetzt, sie wird mit geeigneten Verfahren der Computerlinguistik aufbereitet.

Von Michael Strube

In Kürze

- Für Zwecke der **Künstlichen Intelligenz** mangelt es den Computern immer noch an Wissen über (für Menschen) selbstverständliche Dinge.
- Dieses **Weltwissen** manuell zusammenzutragen ist überaus mühsam, die automatische Gewinnung aus großen Textmengen ist fehlerträchtig.
- Aus der Online-Enzyklopädie **Wikipedia** lässt sich Weltwissen mit weit gehend automatisierten Verfahren extrahieren.
- Die Ergebnisse sind bereits zum Abgleich von Wikipedia-Artikeln in **verschiedenen Sprachen** einsetzbar.

Computer können viel: Aufgaben, für die ein Mensch seine ganze Intelligenz und viel Zeit einsetzen muss, erledigen sie in Sekundenbruchteilen; sie speichern Mengen an Informationen, die unser Gedächtnis weit überfordern würden, und stellen sie auf Anforderung bereit. Nur mit einigen Dingen, die uns so leicht fallen, dass wir unsere geistige Aktivität kaum bemerken, tun sie sich nach wie vor schwer. Dazu zählt ganz besonders die Aufgabe, natürliche Sprache zu verstehen.

Dem Ziel, diesem Defizit abzuhelfen, widmet sich die Computerlinguistik, ein Zweig des Forschungsgebiets, das allgemein als »Künstliche Intelligenz« (KI) bezeichnet wird. Hier hat es in den letzten Jahren einen großen Sprung nach vorne gegeben, von dem ich im Folgenden berichten will.

Was tun wir, wenn wir einen gewöhnlichen deutschen Satz lesen, und was hindert den Computer daran, es uns gleichzutun? Ein Beispiel:

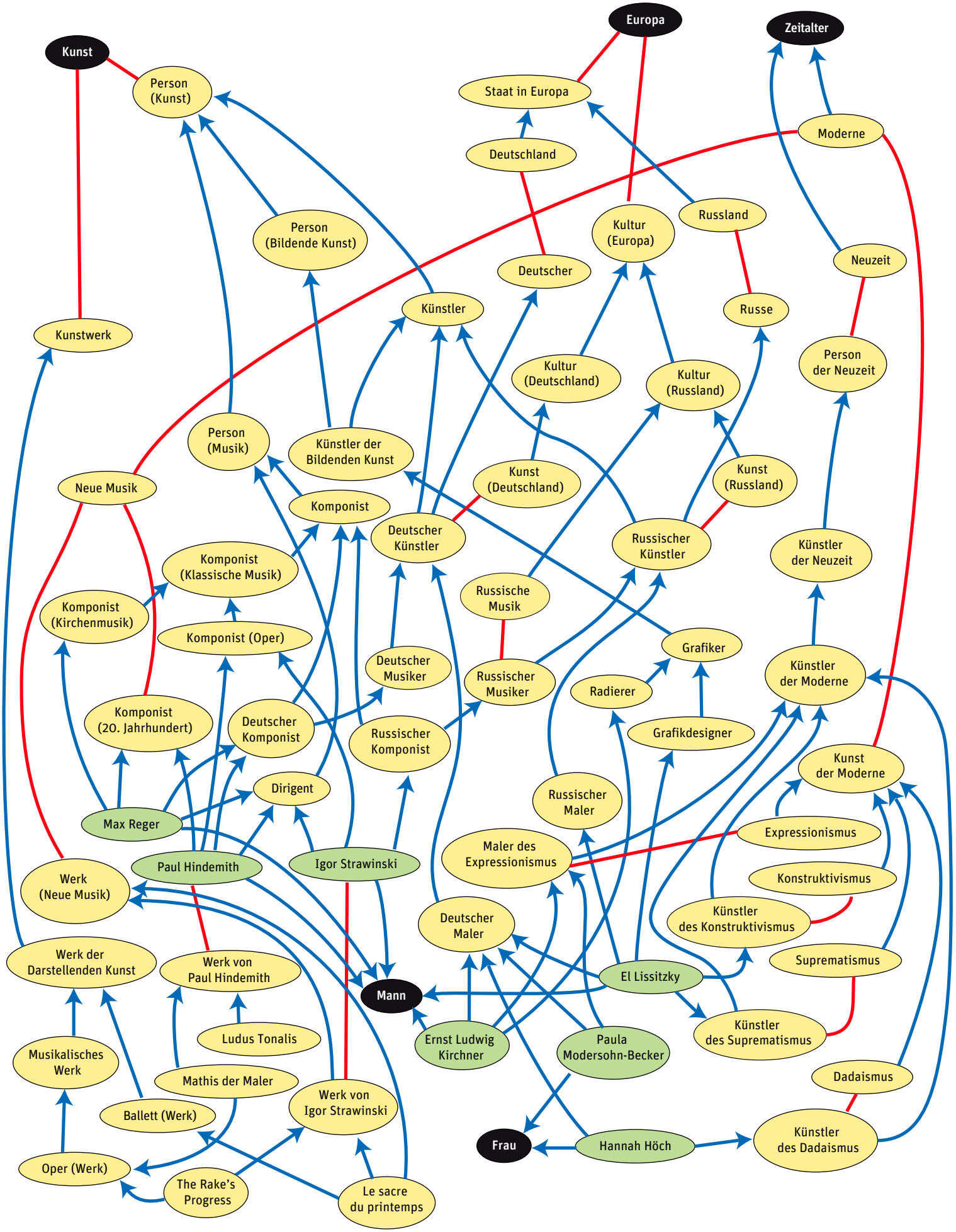
Zu Beginn des Ersten Weltkrieges meldete sich Kirchner als Freiwilliger und wurde Fahrer bei einem Artillerieregiment.

Um diesen Satz zu verstehen, müssen wir erschließen, wer mit Kirchner bezeichnet wird. Dass es sich nicht um die argentinische Präsidentin Cristina Fernández de Kirchner handelt, geht bereits aus der Grammatik des Satzes hervor: Freiwilliger und Fahrer beziehen sich auf eine männliche Person. Es ist nicht besonders schwer, die zugehörigen Regeln in ein Computerprogramm zu fassen. Um dagegen auszuschließen, dass mit Kirchner ihr kürzlich verstorbener Mann, der ehemalige argentinische Präsident Néstor Kirchner, gemeint ist, müssen wir Weltwissen einsetzen, in diesem Fall, dass ein 1950 geborener Mann nicht an einem Ereignis, dass am Anfang des 20. Jahrhunderts stattfand, dem Ersten Weltkrieg, teilgenommen haben kann.

Lesen wir weiter: Im Frühjahr 1915 kam der Künstler nach Halle an der Saale. Wir erfahren nun, dass Kirchner als Künstler bezeichnet wird. Unser Weltwissen verhilft uns zu der Vermutung, dass es sich um den Maler Ernst Ludwig Kirchner handelt, der in der Tat zur fraglichen Zeit gelebt hat und als Künstler aktiv war.

An diesem einfachen Beispiel können wir sehen, dass Sprache zu verstehen Wissen er-

Dieser kleine Ausschnitt aus dem Kategoriennetz der deutschen Wikipedia konzentriert sich auf die nähere Umgebung der Einträge zu einigen Künstlern vom Beginn des 20. Jahrhunderts. Relationen vom Typ `isa` (Kasten S. 96) sind blau und mit Pfeil eingezeichnet, solche vom Typ `notisa` rot.


Kunst
Europa
Zeitalter
Person (Kunst)
Staat in Europa
Moderne
Deutschland
Person (Bildende Kunst)
Russland
Kultur (Europa)
Neuzeit
Deutscher
Künstler
Russe
Kunstwerk
Kultur (Deutschland)
Kultur (Russland)
Person der Neuzeit
Person (Musik)
Künstler der Bildenden Kunst
Kunst (Deutschland)
Kunst (Russland)
Neue Musik
Komponist
Deutscher Künstler
Russischer Künstler
Künstler der Neuzeit
Komponist (Klassische Musik)
Russische Musik
Komponist (Kirchenmusik)
Komponist (Oper)
Grafiker
Deutscher Musiker
Russischer Musiker
Künstler der Moderne
Radierer
Komponist (20. Jahrhundert)
Deutscher Komponist
Russischer Komponist
Grafikdesigner
Kunst der Moderne
Dirigent
Russischer Maler
Max Reger
Expressionismus
Maler des Expressionismus
Paul Hindemith
Igor Strawinski
Konstruktivismus
Werk (Neue Musik)
Deutscher Maler
Künstler des Konstruktivismus
Suprematismus
Werk der Darstellenden Kunst
Werk von Paul Hindemith
El Lissitzky
Mann
Künstler des Suprematismus
Ludus Tonalis
Ernst Ludwig Kirchner
Paula Modersohn-Becker
Musikalisches Werk
Dadaismus
Mathis der Maler
Werk von Igor Strawinski
Ballett (Werk)
Künstler des Dadaismus
Frau
Hannah Höch
Oper (Werk)
The Rake's Progress
Le sacre du printemps

Wissensrepräsentation – die Fachausdrücke

Ein Konzept ist grob gesprochen alles, worüber man sinnvollerweise einen Lexikonartikel schreiben kann. Genauer gesagt, verweist ein Konzept

- auf ein konkretes Objekt: Angela Merkel, Wildschwein, Erdbeere, Holzschraube, Künstler, …
- auf ein abstraktes Objekt: Exponentialfunktion, Vektorraum, Gesetz, Sprache, Gattung, …;
- auf einen Zustand: Armut, Ruhe, Gleichgewicht, …;
- auf ein Ereignis: Krieg, Mondlandung, Hochzeit, …

Ein »Objekt« in diesem Sprachgebrauch kann eine Abstraktionsstufe höher liegen als ein Objekt im umgangssprachlichen Sinn. Letzteres ist dann typischerweise eine **Instanz** des ersten: Wildsau Susi aus dem Gehege der Ketscher Rheininsel ist eine Instanz von Wildschwein. Allgemein ist eine Instanz ein Individuum, das heißt ein Objekt, das mit einem Eigennamen bezeichnet wird.

Eine **Kategorie** ist ein Konzept, das Konzepte zusammenfasst, die gemeinsame Eigenschaften haben.

Aussagen werden mit Hilfe von Relationen ausgedrückt. Eine **Relation** ist eine Beziehung zwischen zwei Konzepten; sie darf unspezifisch sein, muss also keine Aussage darüber enthalten, welcher Art diese Beziehung ist. Alle Konzepte einer Wissensbasis zusammen mit ihren Relationen bilden einen mathematischen Graphen namens **semantisches Netz**. Die Knoten des Graphen sind die Konzepte, die Kanten die Relationen.

Unspezifische Relationen können stark oder schwach sein. Die stärkste Relation ist **Synonymie** (die beiden Konzepte sind gleichbedeutend), die schwächste **Assoziation** (die beiden Konzepte haben irgendetwas miteinander zu tun).

Die wichtigste spezifische Relation ist **`isa`** (zu lesen als englisch *is a*, »ist ein«). Sie drückt die Beziehung zwischen Begriff und Oberbegriff aus und ist daher asymmetrisch. Die entsprechende Kante im semantischen Netz trägt gewissermaßen einen Pfeil, der von einem Konzept zu einer ihm übergeordneten Kategorie weist.

Eine **Taxonomie** ist ein semantisches Netz, dessen Kanten sämtlich die Relation `isa` ausdrücken und das als Graph eine Baumstruktur hat, das heißt, alle Wege in Richtung der Pfeile enden an einem ausgezeichneten Knoten, der »Wurzel« des Baums, und es gibt (unter Einhaltung der Pfeilrichtung) keine Rundwege im Graphen.

Eine **Ontologie** ist ein semantisches Netz, dessen Gerüst durch eine Taxonomie gebildet wird, dessen Konzepte aber zusätzlich durch weitere Relationen verbunden sind, etwa `has-part` (»hat als Teil«), `is-located` (»liegt in«), `has-property` (»hat die Eigenschaft«).

fordert, über Objekte in der Welt und deren Relationen zueinander, über Ereignisse, Handlungen, Situationen. Dieses Wissen ist für die Künstliche Intelligenz heute nur sehr bruchstückhaft verfügbar.

Nicht dass es nicht vorhanden wäre: Natürlich kann man beliebig viele deutsche Sätze in einem Computer abspeichern. Aber das Wissen muss in einer solchen Form vorliegen, dass ein Programm daraus automatisch Schlüsse ziehen kann. Auch das Schlussfolgern selbst ist kein großes Problem. Schon die Philosophen der Antike, namentlich Aristoteles, haben dafür formale Regeln aufgestellt, und die mathematische Logik hat im 19. und der ersten Hälfte des 20. Jahrhunderts diesen Formalismus in aller wünschenswerten Tiefe ausgearbeitet. Auf dieser Grundlage konnte die KI bereits in den 1950er Jahren Methoden und Systeme entwickeln, die rationale Schlüsse ziehen können.

Diese Fähigkeit ist jedoch für Intelligenz noch nicht ausreichend, wie John McCarthy, einer der Gründerväter der KI, schon 1959 feststellte. McCarthy beschrieb am Beispiel der Aufgabe, mit dem Auto zum Flughafen zu fahren, welche Arten von Wissen benötigt werden:

- über Objekte in der Welt und ihre Beziehungen zueinander: Zum Auto gehören Motor und Räder;
- über Situationen: Wenn die Ampel rot ist, darf man die Kreuzung nicht überqueren;
- über Ereignisse und Handlungen: Wenn man am Flughafen angekommen ist, muss man einen Parkplatz suchen.

McCarthy forderte darüber hinaus eine klare Trennung zwischen Schlussregeln und der Repräsentation von Wissen.

In der Folge gelang es in den 1970er und 1980er Jahren, aus verfeinerten Schlussverfahren zusammen mit einer überschaubaren Menge von Fakten aus begrenzten Wissensbereichen (»Domänen«) eine Reihe intelligenter Systeme aufzubauen. Allerdings musste das zugehörige Wissen manuell erstellt werden: Ein Mensch hatte jedes einzelne Wissenselement (»ein Auto hat Räder«) in die Datenbank einzutippen, in einer Form, die ein Computerprogramm verarbeiten kann.

Wie bringt man dem Computer bei, was ein Auto ist? In welchem Sinn kann er überhaupt die »Bedeutung« des Wortes Auto erfassen? Typischerweise nur, indem er Beziehungen des Konzepts Auto zu anderen Konzepten wie Rad auswertet. Eine Wissensbasis (ein »semantisches Netz«, siehe Kasten oben) besteht also aus (möglichst vielen) Konzepten und Relationen zwischen ihnen. Auf dieser Grundidee beruht auch der groß angelegte Versuch, über das ganze Internet ein semantisches Netz, das *semantic web*, zu spannen (Spek-

trum der Wissenschaft 8/2001, S. 42, und 11/2008, S. 92).

Für große Domänen ist das manuelle Anlegen einer Wissensbasis nicht nur überaus aufwändig, sondern stößt auch an prinzipielle Grenzen. Eine Formalisierung, die für ein begrenztes Gebiet, sagen wir die Systematik der Blütenpflanzen, gut funktioniert, lässt sich in der Regel nicht über ihr Gebiet hinaus verallgemeinern. Aus demselben Grund stieß auch der Versuch, verschiedene Domänenwissensbasen miteinander zu verknüpfen, auf unüberwindliche Schwierigkeiten.

Handverlesenes Wissen: Cyc und WordNet

Nur zwei Großprojekte haben diesen Schwierigkeiten zum Trotz überlebt: Cyc und WordNet. Beide stammen aus den 1980er Jahren. Cyc, entworfen von dem KI-Forscher Douglas Lenat in Austin (Texas), stellt neben Konzepten und Relationen auch umfangreiche Software als Schnittstelle zu sprachverarbeitenden Systemen zur Verfügung und bringt einen eigenen Mechanismus zum logischen Schließen mit. WordNet, ins Leben gerufen durch den Kognitionspsychologen George A. Miller in Princeton (New Jersey), verwendet Erkenntnisse der Kognitionspsychologie und statistische Analysen von Texten. WordNet ist eine der am meisten genutzten Ressourcen in der automatischen Sprachverarbeitung und liegt im Gegensatz zu Cyc mittlerweile für viele Sprachen vor, für das Deutsche als GermaNet.

Wozu verwendet man das viele Wissen, das in einer großen Wissensbasis dieser Art niedergelegt ist? Eine wesentliche Anwendung ist die Textanalyse. Ein entsprechendes Programm soll einem vorgelegten Text ungefähr die gleiche Interpretation zuweisen wie ein Mensch. Insbesondere soll es einen gesprochenen Text korrekt niederschreiben, Wortarten erkennen (»Ist dieses Wort ein Substantiv oder ein Adjektiv?«) und die Grammatik eines geschriebenen Textes erfassen (»Wo beginnt der Nebensatz?«, »Was ist das Subjekt?«). Wissen kommt ins Spiel beim Erkennen von Eigennamen, dem Auflösen von mehrdeutigen Wörtern, der so genannten Disambiguierung (»Ist mit Bank ein Sitzmöbel oder ein Kreditinstitut gemeint?«). Eine der zentralen Aufgaben bei der Textanalyse, für deren Lösung Wissen benötigt wird, ist die Koreferenzresolution: Auf welchen bereits genannten Begriff bezieht sich dieses Pronomen? Oder allgemeiner: Welche (verschiedenen) Ausdrücke im Dokument, zum Beispiel Pronomina und umschreibende Phrasen, verweisen auf dasselbe Objekt in der Welt?

Sowohl Cyc als auch WordNet werden seit etwa 25 Jahren kontinuierlich betrieben und setzen qualitätssichernde Maßnahmen ein. Dennoch haben sie weder quantitativ noch qualitativ einen zufrieden stellenden Stand erreicht. Beide Wissensbasen enthalten Bereiche, die vor mehr als einem Jahrzehnt modelliert wurden. Niemand weiß, ob dieses Wissen noch aktuell ist und sich mit später hinzugefügtem Wissen verträgt. Teilbereiche beider Wissensbasen wurden von Einzelpersonen modelliert. Es ist nicht klar, ob ein anderer Modellierer das Wissen auf die gleiche Art strukturiert hätte. WordNet weist (absichtlich) große Lücken bei Instanzen auf, also Konzepten, die durch Eigennamen bezeichnet werden, und muss deshalb in diesem Bereich durch automatische Verfahren ergänzt werden. So findet man zwar, wenn man nach Kirchner sucht, den Eintrag über Ernst Ludwig Kirchner. Die Suche nach der aktuellen Präsidentin Argentiniens bleibt jedoch ohne Erfolg. Cyc hat in diesem Bereich mehr zu bieten, trifft jedoch unter den Instanzen eine recht willkürliche Auswahl.

Zu Beginn der 1990er Jahre kam der manuelle Aufbau von Wissensquellen aus der Mode. Statistische Verfahren hielten Einzug in die Computerlinguistik, nachdem sich herausstellte, dass mit ihren Mitteln die einfacheren Fragen der Textanalyse viel effizienter zu beantworten waren als durch die klassische, auf Schlussverfahren beruhende Analyse. Auch die automatischen Übersetzungsprogramme kommerzieller Suchmaschinen arbeiten mit Material, das durch statistische Auswertung sehr großer Datenmengen gewonnen wurde.

Erst in jüngster Zeit werden die Grenzen der statistischen Verfahren immer deutlicher sichtbar, und man besinnt sich auf linguistisches und Weltwissen zurück. Allerdings ist der manuelle Aufbau von Wissensbasen keine realistische Option mehr, schon wegen der extrem hohen Kosten. Der Aufbau von Cyc soll weit mehr als 1000 Personenjahre in Anspruch genommen haben. Inzwischen war jedoch das World Wide Web zu gewaltiger Größe herangewachsen. Da keimte bei den KI-Forschern die Hoffnung, man könnte aus den dort bequem verfügbaren großen Textmengen automatisch und deshalb kostengünstig Wissen extrahieren.

Das daraufhin entwickelte Verfahren heißt »Ontologielernen aus Texten« und verwendet in erster Linie sprachliche Muster, das sind Textschablonen mit Leerstellen, in die beliebige Wörter oder Wortteile eingesetzt werden können. Ein Beispiel ist das Muster * wie * und * (die Sterne stehen für die Leerstellen).

Es ist überaus mühsam, Weltwissen, das jeder kennt, manuell in eine Datenbank einzuspeisen

Ontologielernen aus Texten nimmt die Formulierung **Anwälte und andere Katastrophen** wörtlich – und hält daraufhin einen Anwalt für so etwas wie ein Erdbeben

Ein Programm, das mit diesem Muster Text durchsucht, findet zum Beispiel die Wortfolge Expressionisten wie Kirchner und Nolde. Daraus kann es ableiten, dass Kirchner und Nolde Expressionisten sind. Mit Hilfe des Musters * und andere * wird die Wortfolge Kirchner und andere Expressionisten gefunden; auch sie erlaubt den Schluss, dass Kirchner ein Expressionist ist. Durch * des *s findet man etwa Werke des Zeichners und daraus Relationen der Art »A gehört zu B«, »B besitzt A« oder »B hat A geschaffen«. Diese Muster kann man ohne großen Aufwand für viele Sprachen erstellen oder sogar ein Programm schreiben, das diese automatisch lernt. Daher ist der Ansatz auch für Sprachen geeignet, für die es nur wenige oder gar keine computerlinguistischen Vorarbeiten gibt.

Allerdings ist das Ontologielernen aus Texten der sprachlichen Kreativität der Schreiber relativ schutzlos ausgesetzt. Da findet das Programm neben Zecken und andere Blutsauger auch die Wortfolge Verwertungsgesellschaften und andere Blutsauger und erschließt daraus, dass Zecken und Verwertungsgesellschaften semantisch ähnlich und Unterbegriffe von Blutsauger sind. Das gleiche gilt für Erdbeben und andere Katastrophen und Anwälte und andere Katastrophen, woraufhin das System Anwälte ebenso wie Erdbeben unter Katastrophen einsortiert. Die Forscher verwenden große Mühe darauf, die Genauigkeit ihrer Ergebnisse zu erhöhen – leider mit mäßigem Erfolg, da die Textquellen auf der einen Seite zu viel irrelevante Informationen enthalten, auf der anderen Seite zu wenig Information, die für KI-Anwendungen relevant wäre.

Wir kennen also bisher zwei Wege, an die Schätze des Weltwissens zu kommen: Man fertigt sie einzeln in mühsamer Handarbeit und bekommt nie genug zusammen; oder man fischt sie mit Netzen aus dem großen Datenstrom und hat mit großen Mengen Unrat bei magerer Ausbeute zu kämpfen. Gibt es einen Mittelweg zwischen beiden Extremen? Gibt es einen Datenstrom, der einerseits handgefertigt und gut strukturiert, andererseits groß genug und ohne zusätzlichen Kostenaufwand verfügbar ist? Die Antwort ist ja. Es handelt sich um die Wikipedia, die seit 2001 bestehende und seitdem ins Gigantische wachsende Internet-Enzyklopädie.

Ordnung durch Kategorien

Ein genauer Blick zeigt, dass die Wikipedia bereits ein semantisches Netz in geeignet formalisierter Gestalt enthält, nämlich das Netz ihrer Kategorien. Fast jeder Artikel ist in eine oder mehrere Kategorien eingeordnet; jede Kategorie ist wiederum selbst kategorisiert. Zum Beispiel ist Ernst Ludwig Kirchner kategorisiert als Maler des Expressionismus, Radierer, Holzschneider, Künstler (documenta), Berliner Secession, Deutscher Maler, Künstler (Aschaffenburg), Geboren 1880, Gestorben 1938, Mann. Das Bild auf S. 95 zeigt einen kleinen Ausschnitt des aus den Kategorien gebildeten semantischen Netzes.

Ein solcher Kategorien-Eintrag stellt eine Relation her, zum Beispiel zwischen den Konzepten Ernst Ludwig Kirchner und Maler des Expressionismus. In diesem Fall sagt die Relation »Ernst Ludwig Kirchner ist ein Maler des Expressionismus«, ist also die Relation `isa` (»ist ein«), die Begriff und Oberbegriff verbindet (Kasten S. 96).

In anderen Fällen, wie zum Beispiel Künstler (Aschaffenburg), versammelt eine Kategorie nur Konzepte, die in irgendeinem Zusammenhang zueinander stehen. Wikipedia-Kategorien sollen in erster Linie dem Leser helfen, verwandte Informationen und Themen leicht zu finden. Die schlichte Tatsache, dass ein Artikel einer Kategorie zugewiesen ist, sagt also noch nicht viel.

Trotzdem kann man aus der Gesamtheit der durch Kategorien definierten Relationen bereits Schlüsse auf die Verwandtschaft zweier Konzepte ziehen. Das haben mein Kollege Simone Paolo Ponzetto und ich 2006 aufgezeigt. Nach Analyse des Wikipedia-Kategoriennetzes konnte unser Programm Fragen der Art »Wie verwandt sind die Konzepte Ernst Ludwig Kirchner und Igor Strawinsky?« (oder »Wie verwandt sind Suprematismus und russische Musik?«) mit einer Maßzahl zwischen 1 (»nicht verwandt«) und 10 (»gleichbedeutend«) beantworten. Überraschenderweise traf es dabei die semantische Verwandtschaft der entsprechenden Konzepte, wie sie durch Befragung menschlicher Versuchspersonen er-

mittelt wird, ungefähr so gut oder sogar besser als das viel aufwändiger erstellte WordNet.

Es gibt allerdings nur relativ wenige Wortpaare, deren Verwandtschaftsgrad durch Befragung von Versuchspersonen ermittelt worden ist. Entsprechend schmal ist die Vergleichsbasis. Daher haben wir unser semantisches Netz noch einem härteren Test unterzogen: Wir machten es zur Arbeitsgrundlage eines Textanalysesystems. Um einen längeren Text in dem oben genannten Sinn zu verstehen, muss ein solches Programm häufig mehrere 100 000 Wortpaare auf ihren Verwandtschaftsgrad testen.

Hierbei zeigte sich, dass das aus Wikipedia abgeleitete Wissen nicht nur qualitativ hochwertig ist, sondern auch eine größere Abdeckung aufweist als WordNet. Da WordNet die meisten Eigennamen nicht kennt, ist unser Programm bei derartigen Konzepten schon deshalb besser, weil es überhaupt eine Antwort liefert. Dieser Vorsprung wird sich noch vergrößern, weil die Wikipedia gegenwärtig vorwiegend in diesem Bereich wächst. Heutige Textanalysesysteme werden in erster Linie auf Zeitungstexte angewendet, und dafür ist es wichtiger, über die argentinische Präsidentin Cristina Fernández de Kirchner Bescheid zu wissen als über den Maler Ernst Ludwig Kirchner.

Taxonomie mit 100 000 Relationen

Noch wesentlich mehr Ausdruckskraft – und damit Nützlichkeit für KI-Anwendungen – gewinnt man, wenn man nicht, wie oben dargestellt, alle Relationen gleichbehandelt, sondern sie nach ihrer Art unterscheidet. Dabei kommt es vor allem darauf an, zu erkennen, ob eine Kategorie die Relation `isa`, die Beziehung zwischen Begriff und Obergriff (Kasten S. 96) ausdrückt. Denn mit deren Hilfe lässt sich eine Taxonomie aufstellen, das heißt eine hierarchische Einteilung der Konzepte, ähnlich wie die Biologen die Arten nach Gattungen, Familien, Ordnungen und so weiter zu gruppieren pflegen. Eine Taxonomie wiederum erlaubt viele Schlussfolgerungen.

Die englische Wikipedia im Zustand vom 25. September 2006, mit der wir gearbeitet haben, enthält 127 325 Kategorien, nicht gerechnet diejenigen, die Wartungszwecken dienen. Da ist an eine manuelle Sortierung der zugehörigen Relationen in `isa` und andere nicht zu denken. Vielmehr haben Ponzetto und ich ein Verfahren entwickelt, das automatisch in mehreren Schritten die Relationen in `isa`, `notisa` (das heißt sicher nicht `isa`) und einen unklassifizierten Rest aufteilt.

Im ersten Schritt verarbeitet das Verfahren den Namen der Kategorie. Viele dieser Namen haben den syntaktischen Kopf gemeinsam: Künstler des Suprematismus ist als Künstler der Moderne kategorisiert, Künstler der Moderne (über einige Zwischenschritte) als Künstler. Alle diese Kategorien haben den gleichen syntaktischen Kopf Künstler, deshalb kann geschlossen werden: Künstler des Suprematismus `isa` Künstler der Moderne und Künstler der Moderne `isa` Künstler. Wenn dagegen ein Wort einmal als syntaktischer Kopf, das andere Mal aber an anderer Position erscheint, können die beiden Konzepte nicht durch isa verbunden sein: Maler des Expressionismus `notisa` Expressionismus.

Wenn der Name der Kategorie im Plural steht (Präsidenten der USA), dann spricht viel dafür, dass die Kategorie alle Träger dieser Eigenschaft aufzählt. Mit einigen Zusatzbedingungen lässt sich die zugehörige Relation mit großer Genauigkeit als `isa` erkennen.

Im zweiten Schritt erzeugt das Verfahren aus bekannten `isa`-Relationen neue. Aus Ernst Ludwig Kirchner `isa` Künstler der Moderne und Künstler der Moderne `isa` Künstler folgt Ernst Ludwig Kirchner `isa` Künstler.

Im dritten Schritt versucht das Verfahren die verbleibenden Relationen mit Hilfe sprachlicher Muster in `isa` und `notisa` zu klassifizieren. Das verläuft so wie beim oben beschriebenen Ontologielernen aus Texten, aber ohne dessen Schwächen, weil schon vor dem Test bekannt ist, dass die beiden Konzepte in einer semantischen Relation stehen. Die Frage, ob ein Anwalt eine Katastrophe ist, wird gar nicht erst gestellt.

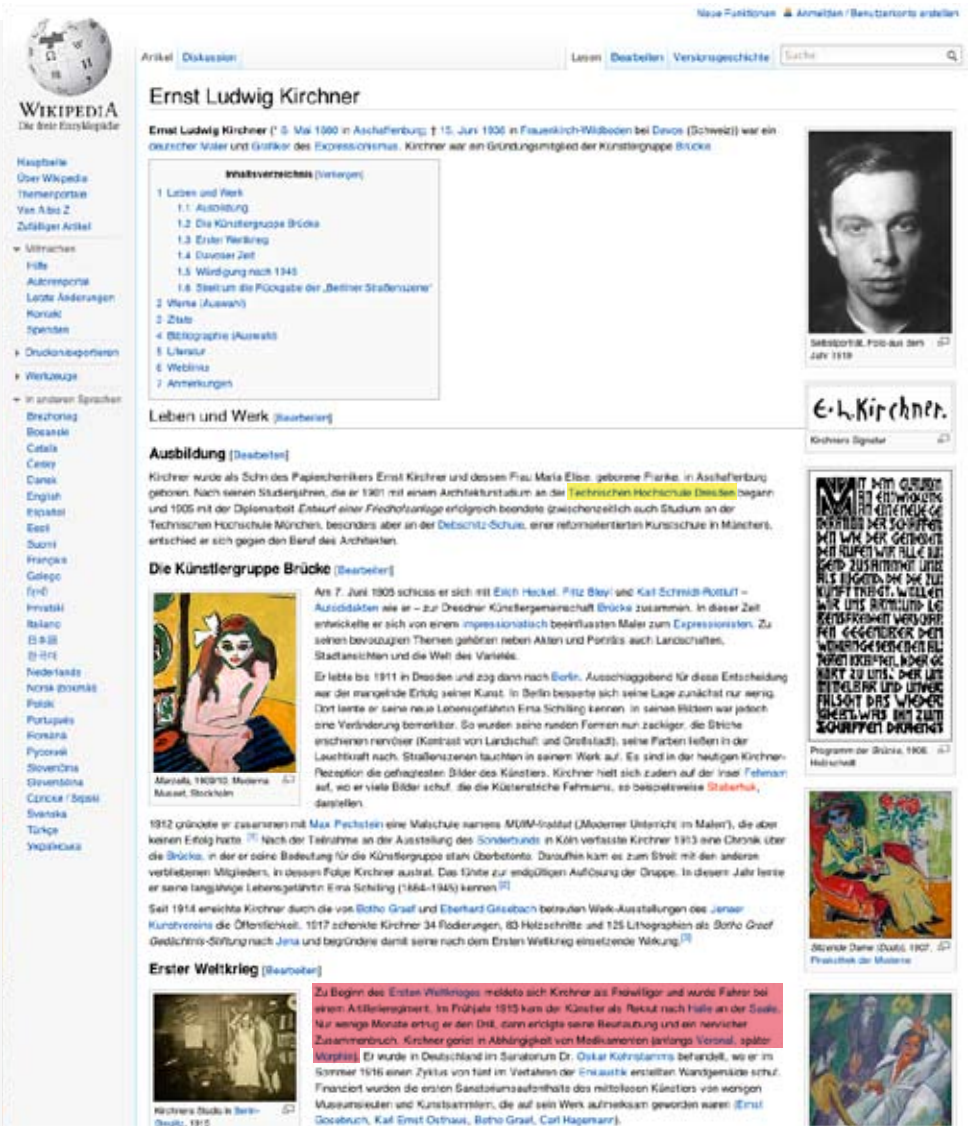
Insgesamt hat unser Verfahren anhand der Wikipedia reichlich 100 000 `isa`-Relationen ausfindig gemacht. Die dadurch induzierte

Textverwandtschaft

Bibliothekssysteme, Fachinformationsdienste und Internet-Suchmaschinen haben, um die Anforderungen ihrer Nutzer zu erfüllen, millionenfach die Frage zu beantworten: »Wie (bedeutungs)ähnlich sind sich zwei vorgelegte Texte?«

Ein automatisches Verfahren liefert eine gute Näherung an eine Antwort. Im Prinzip zählt man aus, wie häufig welche Wörter in beiden Texten vorkommen, multipliziert die Häufigkeit eines Worts in Text *A* mit der Häufigkeit desselben Worts in Text *B* und addiert alle Produkte zusammen (Skalarprodukt der beiden Vektoren, die den Texten entsprechen). Je größer das Ergebnis, desto näher sind die beiden Texte verwandt. Das Prinzip muss noch verfeinert werden: Je exotischer ein Wort ist, das heißt, je seltener es im ganzen Textkorpus vorkommt, desto mehr zählt es mit (sonst würden Allerweltswörter wie »der«, »es« oder »und« eine Rolle spielen); und statt der absoluten Häufigkeiten sind relative zu verwenden (man teile durch die Textlänge).

Das Ergebnis lässt sich geometrisch als Winkel in einem abstrakten Raum interpretieren: Sehr ähnliche Texte haben sehr kleine Winkel (ihre Vektoren zeigen fast in dieselbe Richtung), sehr unterschiedliche Texte stehen praktisch senkrecht aufeinander.



Der deutsche (links) und der französische (rechts) Artikel über Ernst Ludwig Kirchner erzählen dieselbe Lebensgeschichte in sehr unterschiedlicher Form und Ausführlichkeit. Während seine Ausbildungsstätte (gelb unterlegt) im jeweils anderen Text noch durch schlichtes Übersetzen

Taxonomie namens WikiTaxonomy ist in ihrer Leistungsfähigkeit den Wissensbasen WordNet und Cyc vergleichbar; darüber hinaus ergänzt sie diese genau dort, wo sie ihre großen, nur durch sehr mühsame Handarbeit zu stopfenden Lücken haben.

Weitere, reichhaltigere Relationen konnten wir ebenfalls durch linguistische Analyse der Kategoriennamen finden. Aus Ernst Ludwig Kirchner `isa` Maler des Expressionismus und Expressionismus `isa` Kunststil kann man Ernst Ludwig Kirchner `kunststil` Expressionismus erschließen, mit der neu definierten Relation `kunststil`, die einem Künstler einen Stil (von möglicherweise mehreren) zuordnet.

Aus El Lissitzky `isa` russischer Maler folgt El Lissitzky `nationalität` Russland. Damit das funktioniert, muss die Analyse die Bestandteile des Expressionismus und russisch als einschränkende Attribute erkannt haben. Das gelingt mit Hilfe der Relationen Maler des Expressionismus `isa` Künstler und russischer Maler `isa` Maler.

Ein Verfahren dieser Art haben wir auf die gesamte Wikipedia angewandt. Das Ergebnis ist ein dicht geknüpftes Netz namens WikiNet aus einfachen, grundlegenden Konzepten und einem reichhaltigen Inventar semantischer Relationen. Die häufigsten Relationen sind neben `isa` solche, die räumliche Beziehungen beschreiben; es folgen Relationen zur Nationalität der beschriebenen Personen sowie zu Gegenstand und Genre eines künstlerischen Werks.

Alternativ zu unserem Ansatz kann man die aus Wikipedia abgeleiteten Konzepte und Teilwissensbasen in die bereits existierende Taxonomie einer Wissensbasis wie WordNet einhängen. Damit ergänzen sich beide Quellen in idealer Weise: WordNet liefert die sorgfältig manuell aufgebaute Grundstruktur, die der Wikipedia fehlt, und Wikipedia das Wissen zu aktuellen Ereignissen und zu Eigennamen, die dem WordNet fehlen. Gjergji Kasneci und seine Kollegen in der Arbeitsgruppe von Gerhard Weikum am Max-Planck-Institut für Informatik in Saarbrücken haben diese Idee erfolgreich realisiert.

Bereits aus dem Kategoriensystem der Wikipedia lässt sich also eine Fülle an maschinenverwertbarem Wissen extrahieren. Darüber hinaus enthält die Wikipedia eine große Menge nutzbaren Materials:

➤ **Infoboxen:** Viele Artikel enthalten eine so genannte Infobox, die standardisierte Information über den Gegenstand enthält. So enthält jeder Artikel über eine deutsche Stadt Felder für Bundesland, Regierungsbezirk, Höhe über dem Meeresspiegel, Fläche, Einwohnerzahl und mehr. Viele Artikel über Personen enthalten eine Infobox, die je nach Beruf unterschiedlich gestaltet ist.

Die Infoboxen enthalten Information so strukturiert wie in einer Datenbank. Diese sind leicht in das Schema Konzept – Relation – Konzept zu überführen; für deutsche Städte wären dann die Relationen `liegt-in-Bundesland`, `liegt-in-Regierungsbezirk`, `hat-Höhe` und so weiter zu definieren. Die so aus den Infoboxen extrahierte Information ist besser strukturiert als die im Kategoriennetz enthaltene, allerdings weit weniger umfangreich. Jedenfalls kann die eine Sorte Informa-

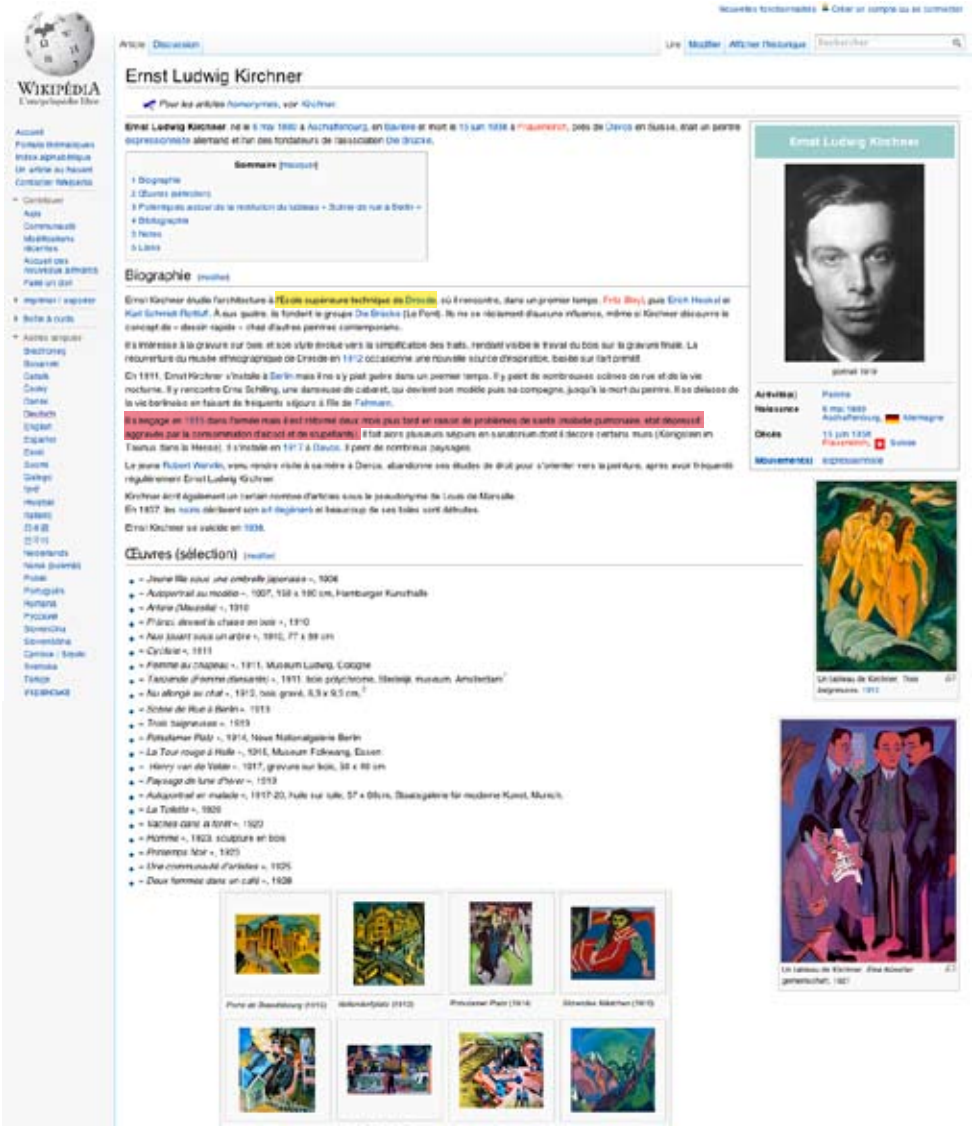
wiederzufinden ist, schreiben beide Artikel über Kirchners Schicksal im Ersten Weltkrieg (rot unterlegt) mit völlig verschiedenen Worten. Gleichwohl kann die in diesem Artikel beschriebene semantische Analyse die beiden Textabschnitte einander zuordnen.

tion dazu dienen, die andere zu überprüfen und gegebenenfalls nachzubessern.

➤ **Hyperlinks:** Wikipedia-Artikel sind in ihrem Text reichlich mit Verweisen auf andere Artikel ausgestattet. Das Netz dieser Hyperlinks kann man auf dieselbe Weise analysieren, wie das beispielsweise die Suchmaschine Google mit ihrem PageRank-Algorithmus für das ganze World Wide Web tut und damit die wichtigen von den unwichtigen Fundstellen unterscheidet. Angewandt auf die Konzepte der Wikipedia arbeitet die Analyse heraus, welche von ihnen zentral und welche peripher sind. Mit dieser Information lässt sich abermals ein semantisches Netz aufbauen. In ihm sind zwei Konzepte durch eine Kante verbunden, wenn irgendeine Assoziation zwischen ihnen besteht. Wieder kann man ihre semantische Verwandtschaft auf einer Skala von 1 bis 10 berechnen und kommt damit zu Ergebnissen vergleichbarer Qualität wie mit dem Kategoriennetz. Allerdings ist auf diesem Weg über die Art der gefundenen Relationen nichts herauszufinden.

➤ **Vielsprachigkeit:** Gegenwärtig liegt Wikipedia in mehr als 250 Sprachen vor. Die größte Version ist die englische mit mehr als drei Millionen Artikeln, gefolgt von der deutschen mit einer reichlichen Million. In zehn Sprachen gibt es mehr als 500 000 Artikel, in 32 Sprachen mehr als 100 000. Alle oben angesprochenen Methoden sind nicht an eine spezielle Sprache gebunden. In der Tat haben Ponzetto und ich semantische Netze für die englische, die deutsche, die französische und die italienische Wikipedia aufgebaut. Die aus den Kategorien abgeleitete Taxonomie liegt sowohl für Englisch als auch für Deutsch vor. Dafür mussten wir lediglich die computerlinguistischen Vorverarbeitungskomponenten austauschen oder anpassen.

Zudem verweisen die meisten Wikipedia-Artikel auf Artikel zum selben Thema in anderen Sprachen. Man könnte das Prinzip dieser *interlanguage links* auf die Taxonomien anwenden mit dem Effekt, dass die Begriffsbäume verschiedener Sprachversionen miteinander verknüpft würden, oder gar alle Sprachen gemeinsam an der Strukturbildung beteiligen. Das letztere Vorgehen ist allerdings so anspruchsvoll, dass es bisher noch nicht realisiert wurde.



Stattdessen haben wir die englische Wikipedia als die größte und am reichhaltigsten strukturierte gewissermaßen als Rückgrat unserer Wissensbasis verwendet. Die anderen Sprachen übernehmen diese Struktur und füllen sie durch Konzepte, wenn und insoweit es Interlanguage-Links zur englischen Wikipedia gibt. Dieses Vorgehen ist praktikabel und bietet eine reichhaltige und einheitliche Struktur für alle Sprachen.

Allerdings gehen gewisse kulturelle Differenzen dabei unter: Für Europäer gehören Kartoffeln, Nudeln und Reis demselben Oberbegriff an, für Asiaten hat Reis eine andere Stellung; für Europäer ist Tomate ein Gemüse, für Asiaten eine Frucht.

➤ **Text:** Die reichste und für den Benutzer wichtigste Wissensquelle der Wikipedia, der Text, ist der linguistischen Analyse am schwersten zugänglich. Evgeniy Gabrilovich und Shaul Markovitch vom Technion in Haifa (Israel) haben jedoch eine Methode gefunden,

auch aus dem Text eines Wikipedia-Artikels in einem automatisierten Verfahren wertvolle Information zu extrahieren.

Ein Standardverfahren der Informationserschließung *(Information Retrieval)* weist Texten einen Verwandtschaftsgrad zu (Kasten S. 99). Gabrilovich und Markovitch trafen nun die Zusatzannahme, dass ein Wikipedia-Artikel eine hinreichend getreue Wiedergabe des zugehörigen Konzepts ist, und schlossen aus der nach dem Verfahren berechneten Bedeutungsverwandtschaft zweier Artikel auf diejenige der zugehörigen Konzepte.

Erkennung der Inhaltsgleichheit

Wie kann die von uns aufgebaute Wissensbasis dazu dienen, in natürlicher Sprache geschriebene Texte zu verstehen? Kommen wir auf das Eingangsbeispiel zurück, das dem deutschen Wikipedia-Artikel über Ernst Ludwig Kirchner entnommen ist.

Zu Beginn des Ersten Weltkrieges meldete sich Kirchner als Freiwilliger und wurde Fahrer bei einem Artillerieregiment. Im Frühjahr 1915 kam der Künstler nach Halle an der Saale. Nur wenige Monate ertrug er den Drill, dann erfolgte seine Beurlaubung und ein nervlicher Zusammenbruch.

Dass sich der Verfasser im zweiten Satz mit Hilfe der Nominalphrase der Künstler auf Kirchner bezieht, kann ein Programm mit Hilfe der aus Wikipedia abgeleiteten Taxonomie überprüfen, denn diese enthält, über Zwischenstufen herleitbar, die Aussage Ernst Ludwig Kirchner `isa` Künstler. Künstler ist also allgemeiner als Kirchner. Dies entspricht dem Aufbau von koreferenten (sich auf denselben Gegenstand beziehenden) Phrasen im Text: Üblicherweise wird zuerst der speziellere Begriff genannt und dann der allgemeinere.

Im dritten Satz werden zwei Pronomina verwendet, zunächst das Personalpronomen er, dann das Possessivpronomen seine. Auch das System kann, wie der lesende Mensch, die Gewissheit gewinnen, dass beide nur auf Ernst Ludwig Kirchner verweisen können. Ernst Ludwig Kirchner `isa` Mann; also ist es zumindest möglich, dass er und seine auf Ernst Ludwig Kirchner verweisen (dass Kirchner und der Künstler identisch sind, weiß das System zu diesem Zeitpunkt schon). Die Möglichkeit wird zur Gewissheit, denn das einzige andere männliche Substantiv, der Erste Weltkrieg, steht erstens viel weiter entfernt, als es für das Bezugswort eines Pronomens üblich ist, und ist zweitens dem System als Ereignis bekannt und damit als etwas, das weder etwas erträgt noch beurlaubt wird.

Eine weitere große Anwendung erarbeiten wir zurzeit im Rahmen des von der Europäischen Union geförderten Projekts CoSyne *(Multilingual Content Synchronization with Wikis)*. Ein Wiki ist ein von vielen Benutzern kooperativ erstellter Inhalt. Häufig existieren Versionen desselben Wikis in mehreren Sprachen nebeneinander her. Prominentestes Beispiel sind die verschiedenen Sprachversionen der Wikipedia; aber auch international agierende Großfirmen dokumentieren ihr internes Wissen, indem alle Beteiligten ohne zentralisierte Arbeitsabläufe in ein und dasselbe Dokument Inhalte einschreiben und dort bearbeiten.

Die dezentrale Arbeitsweise bringt es mit sich, dass häufig die verschiedenen Sprachversionen eines Artikels denselben Gegenstand in sehr unterschiedlicher Weise behandeln. Durch einen automatischen Abgleich dieser Versionen können beide sehr gewinnen, indem Fehlendes eingefügt und möglicherweise Falsches eliminiert wird.

Im Projekt CoSyne sollen Methoden und Computerprogramme entwickelt werden, die erkennen, welche Teile in mehrsprachigen Wikis den gleichen Inhalt haben, feststellen, welche Inhalte in anderen Sprachen fehlen, diese automatisch mit Hilfe eines statistikbasierten Übersetzungsprogramms in die andere Sprache übertragen und an der richtigen Stelle einfügen. Im Allgemeinen wird ein menschlicher Benutzer an diesen Änderungen etwas zu korrigieren haben. Anhand dieser Korrekturen soll das System einen Lernprozess vollführen und so das automatische Übersetzungsprogramm verbessern.

Bereits zur Erkennung der Inhaltsgleichheit ist Wissen erforderlich, das über den Inhalt eines gewöhnlichen Lexikons hinausgeht. Die Wissensbasis kann helfen, Ausdrücke miteinander zu verbinden, die nicht einfach Übersetzungen voneinander, aber gleichwohl semantisch miteinander verwandt sind (Bild S. 100/101). So lässt sich feststellen, welche Abschnitte, Sätze und Teilsätze in welcher Sprache fehlen. Diese Teile soll das System einem statistikbasierten Übersetzungsprogramm anvertrauen und das Ergebnis an der richtigen Stelle einfügen, die es wiederum durch eine strukturelle Analyse der Dokumente bestimmt.

Wir hoffen, dass die Erschließung der Wissensquelle Wikipedia über die genannten Anwendungen hinaus viele weitere Verfahren der Computerlinguistik entscheidend voranbringen kann. Nebenbei haben die hier vorgestellten Arbeiten ein weiteres Mal gezeigt, dass Wikipedia eine Wissensquelle von sehr hoher Qualität ist, die den Vergleich mit manuell erzeugten Wissensquellen und anderen Enzyklopädien nicht zu scheuen braucht. ◁

Michael Strube, Jahrgang 1965, wurde 1996 an der Universität Freiburg mit einer Dissertation in Computerlinguistik promoviert. Nach einer Postdoktorandenzeit an der University of Pennsylvania in Philadelphia kam er 2000 als wissenschaftlicher Mitarbeiter zur EML Research gGmbH in Heidelberg. Ein Jahr später wurde er Leiter der Natural Language Processing Group des Instituts, das mittlerweile »Heidelberger Institut für Theoretische Studien« heißt. Er ist Honorarprofessor an der Universität Heidelberg im Fach Computerlinguistik.

Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, 6 – 12 January 2007. AAAI Press, Menlo Park (California) 2007, S. 1606 – 1611. Online unter www.ijcai.org/papers07/Papers/IJCAI07-259.pdf

Kasneci, G. et al.: The YAGO-NAGA Approach to Knowledge Discovery. In: SIGMOD Record 37(4), S. 41 – 47, 2008.

Ponzetto, S. P., Strube, M.: Knowledge Derived from Wikipedia for Computing Semantic Relatedness. In: Journal of Artificial Intelligence Research 30, S. 181 – 212, 2007. Online unter www.jair.org/media/2308/live-2308-3485-jair.pdf

Ponzetto, S. P., Strube, M.: Deriving a Large Scale Taxonomy from Wikipedia. In: Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence, Vancouver, 22 – 26 July 2007, S. 1440 – 1445, 2007. Online unter www.aaai.org/Papers/AAAI/2007/AAAI07-228.pdf

Weblinks zu diesem Thema finden Sie unter www.spektrum.de/artikel/1050013.