Kreativ durch Analogien

Gleiche Strukturen erkennen bei Dingen, die auf den ersten Blick nichts miteinander gemein haben: Das ist das Arbeitsprinzip, mit dem die interdisziplinäre Computerlinguistik ihre Erfolge erzielt.

Von Michael Strube

Die Computerlinguistik vereinigt Elemente von Informatik und Linguistik; sie verwendet darüber hinaus Methoden aus weiteren Gebieten wie Mathematik, Psychologie, Statistik und künstliche Intelligenz. Der Reiz und die Herausforderung einer solchen interdisziplinären Wissenschaft liegen darin, Analogien zwischen Konzepten aus weit entfernten Teilgebieten zu erkennen und zu nutzen.

Paradebeispiel dafür ist einer der entscheidenden Durchbrüche, welche die Computerlinguistik prägten. Es geht um das »Parsing«: Ein Computerprogramm, genauer gesagt ein Compiler, nimmt Zeichen für Zeichen den Input des Benutzers entgegen, der in diesem Fall seinerseits aus dem Text eines Computerprogramms besteht, und ermittelt dessen Struktur. Im Prinzip dasselbe tut ein Mensch, der einen gesprochenen Satz hört und versteht.

Diese Analogie ist noch nicht besonders bemerkenswert, weil die Entwickler der Programmiersprachen und der zugehörigen Parserprogramme von Anfang an stark von der Linguistik beeinflusst waren; da verwundert es nicht, dass sie deren Denkstrukturen übernommen haben. Aber die Analogie funktioniert auch in Gegenrichtung. Erst als die Informatiker Methoden aus dem Kompilieren formaler Sprachen – insbesondere Programmiersprachen – auf natürliche Sprache übertrugen, wurde das Parsing von gewöhnlichen Sätzen überhaupt effektiv berechenbar. Erst dann konnten sie also Programme schreiben, die einen normalen, gesprochenen Satz hören und in akzeptabler Zeit zumindest seine grammatische Struktur erkennen.

Mehr noch: Ein solches Programm soll vor dem eigentlichen Parsing kontinuierliche Sprache erkennen, das heißt im pausenlosen Strom der gesprochenen Laute einzelne Wörter und damit auch die Grenzen zwischen den Wörtern ausfindig machen, und das unabhängig von der Person des Sprechers und mit großem Wortschatz. Diese Aufgabe in ausreichender Qualität zu lösen, gelang erst mit Hilfe einer weiteren Analogie. Man interpretiert das Sprachsignal als verrauschte, das heißt durch zufällige Störungen verunreinigte Version einer Zeichenkette, die dekodiert werden muss. Dank der neuen Betrachtungsweise lassen sich nun statistische Methoden aus der Informationstheorie anwenden.

Koreferenzresolution mit annotierten Paaren

As we know, Putin has kept putting off this visit to Japan since last year, like back then when Yeltsin repeatedly postponed his trip to Japan.
That is to say, Japan asked for too high a price.
That is, it asked the Russian president to come to Japan to make concessions on territorial issues.
Well, well, the Russian president was still unwilling, was unwilling to make concessions.

Im Text oben sind als koreferent erkannte Erwähnungen farbig unterlegt und durch gleichfarbige Striche miteinander verbunden. Hier kommt es nicht nur darauf an zu verstehen, dass »his« sich auf »Yeltsin« bezieht und »it« auf Japan, sondern auch darauf, dass mit »the Russian president« »Putin« gemeint ist. Letzteres erfordert sogar Weltwissen, nämlich dass zu der Zeit, als dieser Text geäußert wurde, nicht mehr Boris Jelzin, sondern Wladimir Putin russischer Präsident war.

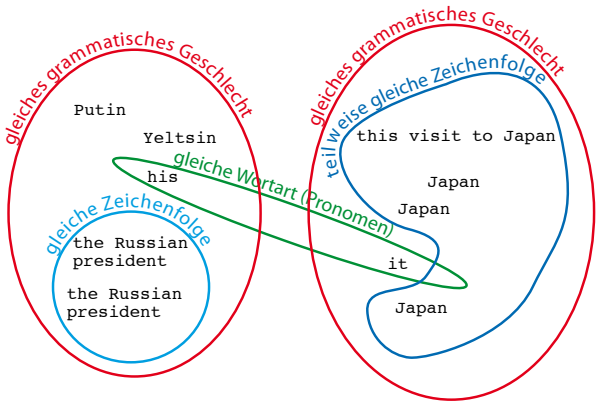
Als Trainingsmaterial für Lernprogramme dienen Listen (»Annotationen«) aus Paaren von Erwähnungen, zum Beispiel aus obigem Text, mit der – von menschlichen Bearbeitern hinzugefügten – Angabe, ob diese Erwähnungen sich auf denselben Gegenstand beziehen (koreferent sind, blauer Strich) oder nicht (roter Strich).

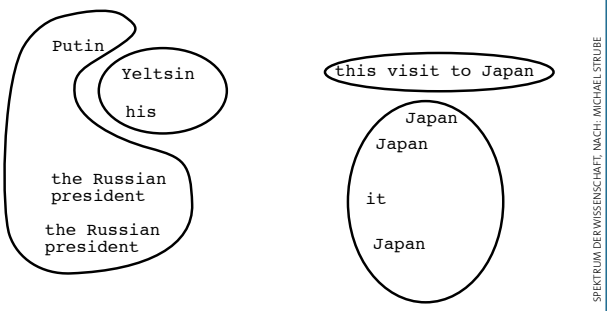
Putin	this visit to Japan	rot
Putin	Japan	rot
this visit to Japan	Japan	rot
Japan	Yeltsin	rot
this visit to Japan	Yeltsin	rot
Putin	Yeltsin	rot
Yeltsin	his	blau
his	Japan	rot
Yeltsin	Japan	rot
Japan	Japan	blau
Japan	it	blau
it	the Russian president	rot
Japan	the Russian president	rot
Japan	the Russian president	rot
his	the Russian president	rot
Yeltsin	the Russian president	rot
Japan	the Russian president	rot
this visit to Japan	the Russian president	rot
Putin	the Russian president	blau
...	...	

SPEKTRUM DER WISSENSCHAFT

Koreferenzresolution mit Hypergraphen

Das Programm definiert zunächst mit Hilfe einzelner Merkmale Teilmengen aller Erwähnungen (Hypergraphen) als Kandidaten für Koreferenzketten. Die Teilmengen sind im linken Bild durch farbige Umrandungen dargestellt. Sie werden dann mit Hilfe von Algorithmen der linearen Algebra verrechnet; das Ergebnis ist die korrekte Zerlegung (rechts).





SPEKTRUM DER WISSENSCHAFT, NACH: MICHAEL STRUBE

Von den so entwickelten Methoden profitierte schließlich immens die maschinelle Übersetzung. Hier trägt dieselbe Analogie: Die Ausgangssprache wird als verrauschte Version der Zielsprache angesehen. Obwohl die automatische Übersetzung auf den ersten Blick nichts mit der Spracherkennung gemein hat, erkannten Computerlinguisten eine Strukturähnlichkeit und übertrugen den Lösungsansatz von der Spracherkennung auf die automatische Übersetzung.

Ist »er« Putin oder Jelzin?

Hier wird ein Muster deutlich: Man löst ein computerlinguistisches Problem, indem man eine Analogie zu einem scheinbar entfernten Gebiet erkennt – natürliche Sprachen und Programmiersprachen, Spracherkennung und Informationstheorie, maschinelle Übersetzung und Spracherkennung. Zwei Studien aus meiner Arbeitsgruppe zeigen im Folgenden, wie eine solche Übertragung im Einzelfall geleistet werden kann.

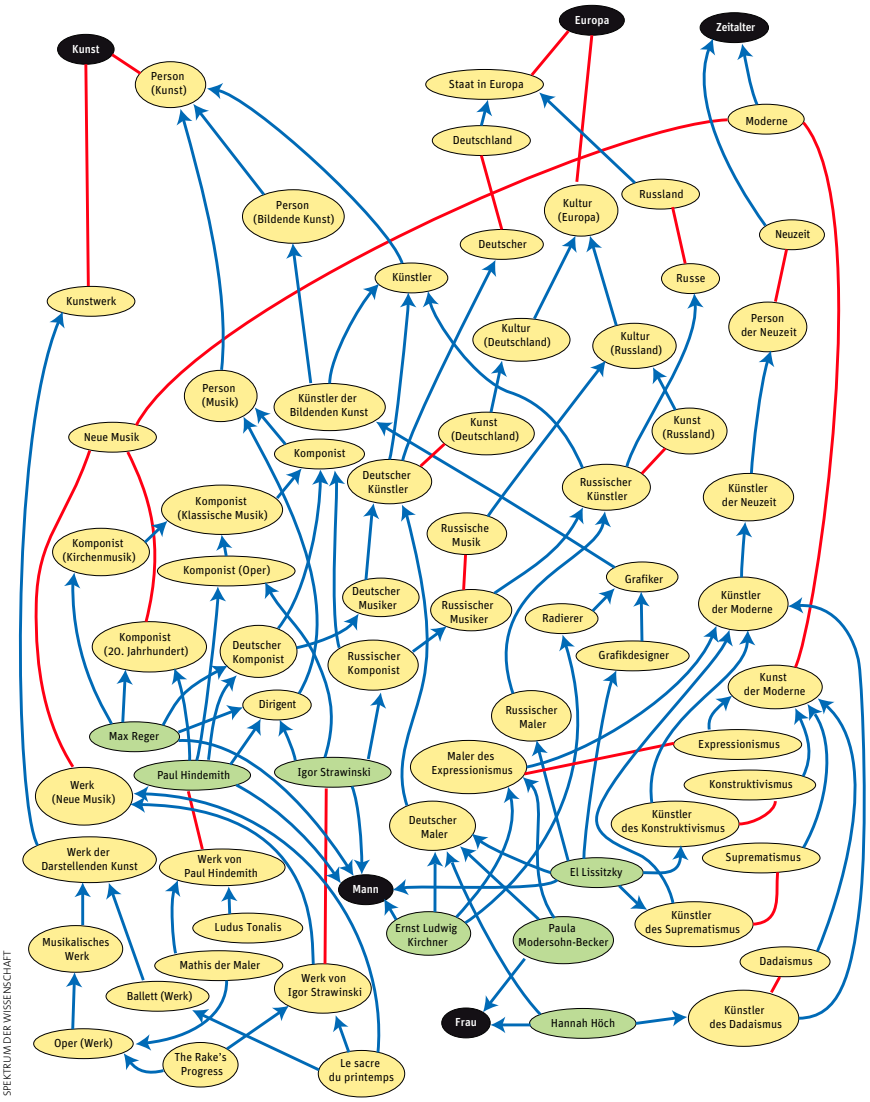
Eine wichtige Aufgabe beim automatischen Verstehen von Texten ist die so genannte Koreferenzresolution: zu erkennen, dass sich mehrere Ausdrücke im Text (»Erwähnungen«) auf denselben Gegenstand beziehen (»koreferieren«). Eine Erwähnung kann zum Beispiel ein Eigenname in unterschiedlichen Varianten, ein Pronomen oder auch eine zusammengesetzte Nominalphrase sein. In dem Text im Kasten links sind die Erwähnungen »Putin« und »the Russian president« koreferent, ebenso »Yeltsin« und »his« sowie »Japan« und »it«. Formal gesprochen kommt es darauf an, alle Erwähnungen in Teilmengen aufzuteilen, deren Elemente zueinander koreferent sind; und natürlich darf eine Erwähnung nicht zwei verschiedenen Teilmengen angehören. Diese Mengen heißen auch »Koreferenzketten«, weil sie häufig, wie im Kasten, durch verbindende Striche dargestellt werden.

Frühe Arbeiten in der Computerlinguistik griffen Erkenntnisse aus der Linguistik auf und stellten komplexe Regeln für die Koreferenzresolution auf, die eine vollständige syntaktische und häufig auch semantische Analyse des Textes voraussetzten. Da dieser Ansatz nicht robust genug für eine Anwendung im größeren Stil war, wurden seit den späten 1990er Jahren zunehmend Verfahren des maschinellen Lernens eingesetzt: Ein Programm leitet automatisch aus von Menschen vorgegebenen korrekten Lösungen, die als Trainings- und Testdaten dienen (»Annotationen«), Regeln oder statistische Zusammenhänge ab.

Damit die Standardverfahren des maschinellen Lernens angewendet werden können, arbeitet man mit Paaren von Erwähnungen. Eine Annotation besteht aus einer Liste solcher Paare mitsamt der Angabe, ob die beiden Erwähnungen eines Paars koreferent sind oder nicht (Kasten links, rechte Grafik). Das Programm lernt nicht nur danach, es gibt zu einem neu vorgelegten Text Listen von Paaren aus. Diese »paarweise Klassifikation« hat den Vorteil, dass sie bekannten und gut verstandenen Methoden des maschinellen Lernens zugänglich ist. Nachteil ist, dass Wissen um den Kontext verloren geht. So kann es einem solchen Programm durchaus passieren, dass es »Putin« und »Yeltsin« durch

Glossar

- **Syntax** ist die grammatische Struktur eines Textes, Semantik seine Bedeutung.
- **Parsing:** Einen Eingabetext Zeichen für Zeichen entgegennehmen, dabei Grenzen zwischen bedeutungstragenden Elementen (»Wörtern«) und in gewissen Grenzen die Struktur des Texts erkennen.
- Zwei Ausdrücke im Text (»Erwähnungen«) **koreferieren**, wenn sie denselben Gegenstand bezeichnen.
- **Koreferenzresolution** ist die Identifizierung koreferenter Erwähnungen.
- **Annotation** ist ein von einem menschlichen Bearbeiter mit Zusatzinformationen versehenes Textbeispiel für das maschinelle Lernen.
- Ein **Synset** ist eine Menge annähernd synonymer Ausdrücke in der Datenbank WordNet.



SPEKTRUM DER WISSENSCHAFT

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. Blaue Pfeile haben die Bedeutung »ist ein« (»ein russischer Komponist ist ein Komponist«, »Igor Strawinski war ein Dirigent«), rote Linien kennzeichnen das Wissen, dass eine solche Beziehung nicht besteht.

eine lange Koreferenzkette verbindet – und damit in einen Topf wirft –, weil an irgendeiner Stelle das Pronomen »his« zum einen wie zum anderen passt und schon die Feststellung, dass »his« nur einen der beiden Herren meinen kann, über die Betrachtung einzelner Paare hinausgeht.

Dies war – stark vereinfacht – der Stand der Forschung, als vor drei Jahren Jie Cai als Doktorandin in meiner Arbeitsgruppe anfing. Wir fragten uns, wie man das Problem der Koreferenz angemessener repräsentieren und insbesondere Wissen über den Kontext in die Entscheidung mit einbeziehen kann. Dabei legten wir das oben beschriebene Konzept zu Grunde, dass Koreferenzketten eigentlich Mengen sind und es darum geht, jede Erwähnung genau einer Menge zuzuweisen. In der Informatik fanden wir ein geeignetes Analogon zu dieser Aufgabe: die Clusteranalyse. Man ordne Datenpunkte Mengen (»Clustern«) zu, und zwar so, dass eng benachbarte Datenpunkte in der Regel in ein und denselben Cluster geraten. Nur kann man zwar zu zwei (durch Koordinaten gegebenen) Datenpunkten in einfacher Weise deren Entfernung definieren; aber das funktioniert für Erwähnungen nicht. Allenfalls sind Erwähnungen Punkte (»Knoten«) in einem Graphen, die genau dann durch eine Kante verbunden sind, wenn sie koreferent sind; aber das ist ja erst das Ergebnis der Analyse und nicht der Ausgangspunkt. Diese Kanten wiederum drücken nichts weiter aus als eine paarweise Klassifikation und bieten daher keinen Fortschritt.

Weiter kommt man mit einem neuen Konzept. Ein Hypergraph ist ein verallgemeinerter Graph, bei dem eine Kante mehr als zwei Knoten miteinander verbinden kann. Damit ist er die graphentheoretische Entsprechung einer Menge, und wir haben eine angemessene Darstellung des Koreferenzproblems gefunden: Erwähnungen sind Knoten im Hypergraphen, und jeder Gegenstand ist eine Hyperkante, die alle seine koreferenten Erwähnungen umfasst. Das Problem der Koreferenzresolution kann dann als Clusteranalyse für Hypergraphen aufgefasst werden.

Mit diesem neuen theoretischen Rahmen ist unser Programm zur Koreferenzresolution nicht mehr ausschließlich auf die Beispielpaare der paarweisen Klassifikation angewiesen. Vielmehr zieht es eine Vielzahl von »Merkmalen« *(features)* heran. Ein Merkmal ist ein Indiz dafür, dass zwei Erwähnungen im Prinzip koreferent sein können. Eines von ihnen zeigt an, ob Erwähnungen der gleichen semantischen Klasse angehören, also zum Beispiel beide eine Person, einen Ort oder ein Fahrzeug bezeichnen. Ein anderes Merkmal stellt dar, ob Erwähnungen die gleiche Zeichenkette enthalten (»Präsident Putin«, »Wladimir Putin«, »Putin«, …). Weitere Merkmale enthalten Wissen über grammatische Eigenschaften einer Erwähnung wie Genus, Numerus und Person, über ihre syntaktische Rolle (Subjekt, Objekt, …) oder bestimmte syntaktische Beziehungen zwischen zwei Erwähnungen, etwa dass eine Erwähnung Apposition einer anderen ist (»Wladimir Putin, der russische Präsident, …«). Auch der Abstand im Text zwischen zwei Erwähnungen, gezählt in Wörtern oder Sätzen, wird als Merkmal ausgedrückt. Insgesamt arbeiten wir mit etwa 20 unterschiedlichen Merkmalen.

Unser Programm erstellt im ersten Schritt zu jedem Merkmal einen Satz von Hyperkanten. Diese sind manchmal gewöhnliche Kanten, verbinden also nur zwei Erwähnungen, zum Beispiel bei dem Merkmal für den Abstand im Text. Die meisten aber umfassen mehr als zwei Erwähnungen; sie machen die Stärke des Verfahrens aus. Allen Hyperkanten werden mit Hilfe von annotierten Trainingsdaten Gewichte zugewiesen; das sind Zahlen,

die bezeichnen, wie stark das mit dem Merkmal ausgedrückte Indiz für Koreferenz ist. Da das Verfahren robust ist gegenüber kleinen Abweichungen bei den Gewichten, kommt es mit fünf Prozent der Trainingsdaten aus, die für die paarweise Klassifikation erforderlich sind. Das ist von entscheidender Bedeutung, da Annotationen für jedes Sachgebiet neu erstellt werden müssen, viele Stunden menschlicher Arbeit erfordern und daher teuer sind.

Die mit Gewichten versehenen Hyperkanten lassen sich in Matrizen umwandeln. Die wiederum kann man mit Standardmethoden aus der linearen Algebra so transformieren, dass am Ende eine korrekte Zerlegung in Mengen koreferenter Erwähnungen steht (Kasten S. 31 oben).

Wikipedia als lexikalische Datenbank

In Experimenten mit Standarddatensätzen konnten Jie Cai und ich zeigen, dass unsere Methode trotz deutlich geringeren Bedarfs an Lernstoff wesentlich bessere Ergebnisse bei der Koreferenzresolution erzielt als die üblichen Verfahren, und das in etwa einem Viertel der Rechenzeit. Wegen des geringen Trainingsaufwands ist es uns auch gelungen, unser Verfahren ohne größere Mühe auf eine neues Sachgebiet zu übertragen: Inzwischen analysiert es nicht nur Nachrichtentexte, sondern auch Arztberichte.

Aufgaben wie die Koreferenzresolution benötigen über das linguistische Wissen (»Ist ›Putin‹ ein Substantiv oder ein Verb?«) hinaus auch Wissen über Objekte in der Welt und ihre Beziehungen zueinander (»Ist ›Putin‹ ein Mensch oder ein Ort?«). Koreferenzrelationen bestehen häufig zwischen einem Unter- und einem Oberbegriff, etwa »Wladimir Putin« und »der russische Präsident«, »der russische Politiker«. Im Oktober 2005 stellte sich meinem damaligen Doktoranden Simone Paolo Ponzetto und mir die Frage, wie wir unserem Koreferenzresolutionssystem dieses Wissen zur Verfügung stellen können.

Die in der Computerlinguistik populärste Ressource für derartiges Wissen ist »WordNet«, eine lexikalische Datenbank, die Wörter so genannten »Synsets« zuordnet, die jeweils eine Menge (annähernd) synonymer Ausdrücke enthalten. Die Synsets sind in einer Taxonomie angeordnet und durch viele weitere semantische Relationen miteinander verknüpft, so dass sich ein reichhaltiges semantisches Netzwerk ergibt. WordNet enthält aber nur wenig Wissen über durch mit Eigennamen bezeichnete Konzepte. So gibt es in der aktuellen Version (Stand 30. Mai 2011) zwar einen Eintrag über »Vladimir Putin«; »Boris Yeltsin« hat allerdings nie Eingang in die Datenbank gefunden. Wir waren also auf der Suche nach einer Wissensquelle, die mehr Informationen über durch Eigennamen bezeichnete Konzepte enthält und dennoch so gut strukturiert ist wie WordNet.

Ein Blick auf die im Oktober 2005 noch recht kleine »Wikipedia« zeigte uns, dass diese Online-Enzyklopädie die erste Bedingung erfüllt. Die zweite Bedingung erforderte einen erneuten, unbefangenen Blick. Im Gegensatz zu gewöhnlichen, unstrukturierten Webseiten enthält Wikipedia neben dem ebenfalls unstrukturierten Text einige Strukturelemente, die unserer Aufgabe dienlich waren. So findet man am Ende jedes Artikels die Liste der Kategorien, denen er angehört. Die Kategorien selbst sind ebenfalls kategorisiert, so dass man mit ihrer Hilfe von einem Artikel zu einem anderen gelangen kann, der mit dem ersten semantisch verwandt ist.

Damit war klar: Wenn es gelingt, aus den Wikipedia-Kategorien ein semantisches Netz zu extrahieren, dann verfügt man über eine Ressource, die WordNet zumindest bei den durch Eigennamen bezeichneten Konzepten überlegen ist. In der Folge haben Ponzetto und ich (später stieß Vivi Nastase als Postdoc zum Team) mehrere Verfahren entwickelt, die Wikipedia zuerst in ein semantisches Netzwerk umwandeln, dann in eine Taxonomie und schließlich in ein Netzwerk mit reichhaltigen semantischen Relationen (Spektrum der Wissenschaft 12/2010, S. 94; Bild S. 33). Die Anwendung auf mehrere computerlinguistische Probleme belegte die Richtigkeit unserer Grundannahme.

Die beiden hier beschriebenen Projekte weisen eine Gemeinsamkeit auf. Beim Problem der Koreferenzresolution kam es darauf an, auf einer abstrakten Ebene die Strukturgleichheit zwischen dem linguistischen Phänomen der Koreferenz, dem mathematischen Konzept der Menge und dem graphentheoretischen Konstrukt des Hypergraphen zu sehen. Bei der Wissensextraktion aus Wikipedia ging es darum, das Kategoriensystem in Wikipedia als Netzwerk zu erkennen, dessen Kanten semantische Nähe ausdrücken und dessen Knoten – Wikipedia-Artikel und -Kategorien – den »Synsets« aus WordNet entsprechen. Hat man diese Strukturgleichheit erst einmal gefunden, ist es relativ leicht, sie zu nutzen – in diesem Fall Wikipedia in ein semantisches Netzwerk umzuformen und darauf weitere Strukturen aufzubauen.

In beiden Beispielen war es entscheidend, Analogien zwischen auf den ersten Blick nicht zusammenhängenden Gebieten zu erkennen. In einem interdisziplinären Gebiet wie der Computerlinguistik gilt dies auch eine Abstraktionsstufe höher: Es kommt darauf an, Analogien zwischen Analogien zu sehen. »Good mathematicians see analogies between theorems or theories. The very best ones see analogies between analogies«, so der bedeutende Mathematiker Stanislaw Ulam (1909–1984) in seinem Werk »Analogies between analogies«.

Die wissenschaftliche Umgebung bei HITS stellt in dieser Beziehung eine einmalige Chance dar, da die Interdisziplinarität zu den Voraussetzungen seiner Existenz zählt. Vielleicht werde ich eines Tages sogar Methoden aus der Biomechanik oder der theoretischen Astrophysik auf computerlinguistische Probleme anwenden!

DER AUTOR

Michael Strube, Jahrgang 1965, wurde 1996 an der Universität Freiburg mit einer Dissertation in Computerlinguistik promoviert. Nach einer Postdoc-Zeit 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.

QUELLEN

Cai, J., Strube, M.: End-to-End Coreference Resolution via Hypergraph Partitioning. In: Proceedings of the 23rd International Conference on Computational Linguistics, Peking, 23.–27. August 2010, S. 143–151. Download über www.aclweb.org/anthology/C/C10/

Ponzetto, S. P., Strube, M.: Taxonomy Induction Based on a Collaboratively Built Knowledge Repository. In: Artificial Intelligence 175, S. 1737–1756, 2011