Verstehen Computer Sprache?

Gerhard Kremer, Institut für Computerlinguistik

Das noch junge Fachgebiet Computerlinguistik ist zwar nur wenigen bekannt, begegnet uns jedoch oft im alltäglichen Leben. Dieses Fach beschäftigt sich mit der Verarbeitung menschlicher Sprache mithilfe des Computers. Methoden der Computerlinguistik sind beispielsweise enthalten in Diktiersystemen (Gesprochenes wird in Text umgewandelt), in Internet-Suchmaschinen (zu einer Suchanfrage werden die relevantesten Internetseiten gesucht), in Telefonauskunftsystemen (gewünschte Informationen werden zusammengefasst und per Computerstimme wiedergegeben) oder in Geräten mit Sprachbedienung (bestimmte Funktionen werden auf gesprochene Befehle hin ausgeführt).

Die Computerlinguistik befasst sich einerseits mit dem strukturellen Aufbau der menschlichen Sprache (Schnittstelle der Computerlinguistik zur Sprachwissenschaft) und im Zusammenhang mit dem Fachgebiet Psycholinguistik mit der Frage „Wie lernen und benutzen Menschen Sprache?", andererseits damit, wie dieses Wissen und diese Erkenntnisse auf dem Computer umsetzbar sind (Schnittstelle der Computerlinguistik zur Informatik). Der in diesem Beitrag vorgestellte Teilbereich der Computerlinguistik, die lexikalische Semantik, beschäftigt sich mit der Bedeutung von Wörtern. Wir werden uns deshalb im Rahmen des Titels auf die Frage konzentrieren: „Wie lernt und versteht der Mensch Wortbedeutung bzw. wie kann das ein Computer?" – und genauer: „Lässt sich (vor allem aus der Sicht des Computers) auf die Bedeutung eines Worts schließen und wenn, dann wie?"

Zur Veranschaulichung eines für Sie unbekannten Worts enthält der folgende Beispielsatz ein Phantasiewort – stellen Sie sich vor, Ihnen reicht jemand ein Paket und sagt:

„Hier bitte, ein Mompel für Sie."

Damit können Sie wahrscheinlich nicht viel anfangen, und Sie werden vielleicht etwas zögernd das Paket entgegennehmen und öffnen, um die Bedeutung des Worts *Mompel* herauszufinden. Wenn jedoch nach diesem Angebot die überbringende Person weiterspricht und Ihnen dadurch mehr Bedeutungszusammenhang (Kontext) bietet – ohne direkt zu sagen, was *Mompel* für ein Ding ist – werden Sie sich trotzdem eine ungefähre Vorstellung von dem Inhalt des Pakets

machen (ihre Reaktionen werden entsprechend unterschiedlich ausfallen). Hier drei Beispiele:

„ … besonders für die Badewanne zu empfehlen."

„ … sieht sicher hervorragend an Ihnen aus."

„ … aber passen Sie auf, er beißt."

Die vorgestellte Situation verdeutlicht einen wesentlichen Aspekt aus der Perspektive eines menschlichen Lerners: Sprachlicher Kontext hilft beim Erschließen der Bedeutung eines Worts. Diese Einsicht bildet die Grundlage einer computerlinguistischen Methode zur Untersuchung der Bedeutungsähnlichkeit von Wörtern und ist bekannt als sogenannte Distributionelle Hypothese: Ein Wort erhält seine Bedeutung durch seine Verwendung zusammen mit anderen Wörtern. Daraus abgeleitet ist die Annahme, dass zwei Wörter, die in ähnlichen Kontexten im Text auftreten, auch eine ähnliche Bedeutung tragen. Wenn wir also in einer Textsammlung (das sogenannte Korpus, das beispielsweise aus Büchern oder Zeitungstexten besteht) zwei Begriffe finden, deren Kontexte oft dieselben Wörter enthalten, dann deutet das nach dieser Annahme darauf hin, dass die beiden Begriffe sich in ihrer Bedeutung ähnlich sind.

Wie wird die Distributionelle Hypothese umgesetzt, damit mit dem Computer Bedeutungsähnlichkeit von Wörtern verarbeitet werden kann? Die Funktionsweise eines Computers beruht schließlich grundsätzlich einfach nur darauf, Zahlen (genauer: Folgen von Nullen und Einsen) zu verarbeiten – die deutsche Bezeichnung ist ‚Rechner'. Also sollte zuerst die Frage beantwortet werden: Wie können wir – zur weiteren Verarbeitung mit dem Computer – die Bedeutung von Wörtern in Zahlen ausdrücken?

Hier ein einfaches Beispiel, um das Vorgehen zu erklären: In einem möglichst großen Korpus suchen wir alle Textausschnitte mit den Zielwörtern, die untersucht werden sollen (z. B. *Bus*). Dann sehen wir uns alle bedeutungstragenden Wörter in der Nähe (dem Kontext) des Zielworts an und zählen, wie oft jedes dieser Kontextwörter mit unserem Zielwort insgesamt vorkam. Angenommen, wir fanden folgende drei Textausschnitte:

[…] *Bus*sen mit *Geschwindigkeit*en zwischen 50 und […]

[…] *Bus* zusammen mit acht neuwertigen *Räd*ern […]

[…] dem *Bus*. Ein *Rad* allein wiegt […]

In diesem Beispiel kommt *Bus* im ersten Textausschnitt zusammen mit *Geschwindigkeit* vor, im zweiten mit *Rad* und im dritten nochmal mit *Rad* (der Einfachheit halber beschränken wir uns hier auf diese beiden Kontextwörter). Angenommen, im Beispiel wäre in den restlichen Texten, die wir zur Verfügung haben, *Bus* noch viermal mit *Geschwindigkeit* und weitere fünfmal mit *Rad* vorgekommen. Wenn wir diese Zahlen in eine Tabelle eintragen, haben wir eine Darstellung des Zielworts, ausgedrückt als ein Paar von Zahlen, die die Vorkommenshäufigkeiten mit den beiden Kontextwörtern angeben – und damit eine gute Verarbeitungsbasis für den Computer bilden. Die erste Zeile in Tabelle 1 fasst zusammen, wie das Zielwort *Bus* durch die Gesamtzahl der gemeinsamen Vorkommen mit den Kontextwörtern *Geschwindigkeit* (an 5 Textstellen) und *Rad* (an 7 Textstellen) dargestellt wird.

	Kontextwörter	
	Geschwindigkeit	Rad
Bus	5	7
Mofa	2	4
Gepard	6	1

Tabelle 1: Gemeinsame Vorkommenshäufigkeiten

Nach der Distributionellen Hypothese sind sich diejenigen Wörter in ihrer Bedeutung ähnlich, die in ähnlichen Kontexten vorkommen. Was sind also ähnliche Kontexte, wenn ein Wort über Zahlen dargestellt ist? Dazu betrachten wir zwei weitere Zielwörter in Tabelle 1, *Mofa* und *Gepard*. Hier ist vorstellbar, dass wir in unseren Texten *Mofa* einige Male im Kontext von *Geschwindigkeit* und *Rad* vorgefunden haben. Plausibel ist auch, wenn *Gepard* vor allem im Zusammenhang mit *Geschwindigkeit* erwähnt wurde, aber selten mit *Rad*. Dadurch unterscheidet sich *Gepard* von *Bus* auch stärker als *Mofa* von *Bus*: Die Verhältnisse der Vorkommenshäufigkeiten sind unähnlicher – der Zusammenhang mit *Geschwindigkeit* bei *Gepard* ist sehr prominent, wobei im Kontext von *Mofa* und *Bus* auch über deren *Geschwindigkeit* geschrieben wurde, aber bei beiden etwa eineinhalb bis zweimal häufiger über *Rad*.

Anschaulicher ist Tabelle 1 in einem Koordinatensystem (wie auf der nächsten Seite abgebildet) dargestellt. Die Vorkommenshäufigkeit des jeweiligen Zielworts mit dem Kontextwort *Geschwindigkeit* kann dort beispielsweise in waagerechter Richtung nach rechts und die gemeinsame Vorkommenshäufigkeit mit *Rad* in senkrechter Richtung nach oben eingetragen werden. Der Punkt für *Gepard* wird also an der Stelle sechs Schritte vom Nullpunkt nach rechts und einen Schritt nach oben markiert. Damit ergibt das Eintragen der Punkte für die drei Zielwörter aus Tabelle 1 eine Darstellung wie in Abbildung 1. Betrachten wir zuerst nur die Lage der drei Punkte zueinander, dann ist hier im Gegensatz zur Tabellendarstellung zuerst nicht so deutlich, dass *Mofa* dem *Bus* nähersteht

als *Gepard* – der Punkt für *Bus* scheint von beiden anderen Punkten fast gleich weit entfernt zu sein.

Alternativ ist zum Berechnen der Bedeutungsähnlichkeit der sogenannte Cosinus-Abstand als Maß üblich. Dabei wird der Nullpunkt mit dem jeweiligen Zielwort-Punkt verbunden und die Richtungen zweier Linien werden verglichen, d. h. der Winkel zwischen diesen beiden Richtungslinien (den sogenannten Vektoren). Je kleiner der Winkel zwischen zwei Vektoren ist, desto ähnlicher sind sich nämlich die Verhältnisse der Vorkommenshäufigkeiten betrachteter Kontextwörter, und desto ähnlicher sind sich nach der Distributionellen Hypothese beide Zielwörter in ihrer Bedeutung. Vergleichen wir den Winkel zwischen den Vektoren für *Bus* und *Mofa* mit dem Winkel zwischen den Vektoren für *Bus* und *Gepard*, wird deutlich, dass *Mofa* hier ähnlicher zu *Bus* ist. Eine praktische Maßzahl für die Ähnlichkeit ist der Cosinus des Winkels, der bei 0 Grad (exakt gleiche Richtung; höchste Ähnlichkeit) den Wert 1 hat, mit zunehmendem Winkel immer kleiner wird, und bei 90 Grad den Wert 0 erreicht. Dieser Cosinus-Abstand ist auch ohne Schaubild direkt aus den Vorkommenshäufigkeiten berechenbar.

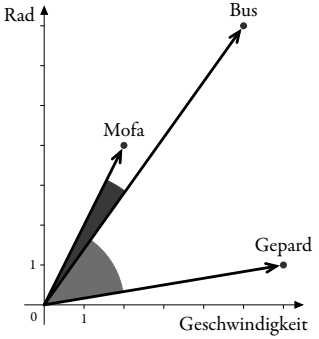

Abbildung 1: Gemeinsame Vorkommenshäufigkeiten im Koordinatensystem

Ein praktischer Grund für den Vergleich von Richtungen der Vektoren ist die Situation, wenn in einer speziellen Textsammlung beispielsweise viel mehr über den *Bus* geschrieben wird als über das *Mofa*. Wenn beide in ähnlichen Kontexten im Text vorkommen, sind die Zahlen für die Vorkommenshäufigkeiten zwar in ähnlichen Verhältnissen, aber sie sind bei *Bus* alle höher als bei *Mofa* – dadurch wird der Vektor im Schaubild also länger (der Punkt ist weiter rechts und weiter oben), aber die Richtung bleibt fast gleich.

Die tatsächliche Anwendung der hier am anschaulichen Beispiel gezeigten Methode erfordert einige Änderungen: Als Kontextwörter werden alle gefundenen Wörter in einem Kontextbereich von wenigen Wörtern bis zu vielen Sätzen einbezogen (außer Nomen auch Verben, Adjektive und Adverbien), und als Zielwörter werden alle vorkommenden Wörter in der Textsammlung benutzt. Mit mehr als drei Kontextwörtern wird eine bildliche Vorstellung eines Schaubilds zwar schwerer, die Berechnung des Cosinus-Abstands funktioniert aller-

dings auch dann noch nach dem gleichen Rechenschema. Gleichzeitig enthält die Tabellendarstellung der Zielwörter dadurch sehr viele Einträge (aber die meisten davon sind nur Nullen und kleine Zahlen und dadurch wenig relevant), die Speicherplatz verbrauchen und rechenzeitaufwändig sind und deshalb einen intelligenten Umgang damit erfordern. Die untersuchten Wörter stehen wie in den Beispiel-Textausschnitten oft nicht in ihrer Grundform im Fließtext – für die Rückführung auf die Grundform (z. B. von „Rädern" auf „Rad") sind weitere Methoden der Computerlinguistik als Vorverarbeitungsschritt nötig, um die großen Datenmengen möglichst zuverlässig und schnell aufzubereiten. Die Qualität der Ergebnisse dieser Methode wird auch durch die Größe und Auswahl der Textsammlung bestimmt. Bei der Nutzung einer größeren Menge von Texten ist die Wahrscheinlichkeit höher, dass mehr Zielwörter in den Kontexten gefunden werden, die relevant für die Bedeutung des Worts sind. Wenn der Fokus des Projekts ein spezielles Themengebiet ist, sollten passende Texte über dieses Thema dafür ausgewählt werden; für allgemeine Anwendungen ist eine gute Mischung aus verschiedenen Textsorten (Zeitungen, Bücher, Konversationen) zu finden.

Eine Schwierigkeit der vorgestellten Methode sind mehrdeutige Wörter (z. B. *Blatt* aus Papier oder als Teil eines Baums), deren unterschiedlichen Bedeutungen in einer gemeinsamen Darstellung vermischt werden, weil alle Kontextwörter (z. B. *schreiben*, *Seite*, *Nest* und *Wind*) in die Darstellung des (gleichlautenden) Zielworts einfließen. So enthält das Zielwort *Blatt* im Ergebnis Teile beider Bedeutungen (da es insgesamt mit allen obigen Kontextwörtern zusammen gefunden wurde), ist also teilweise ähnlich zu *Papier* und teilweise zu *Baum*. Dadurch hat *Blatt* zu keinem der beiden Wörter besonders starke Ähnlichkeit, da in einem Vergleich die Anteile beider Bedeutungen in die Berechnung einbezogen werden.

Die Unterscheidung der Art der Bedeutungsähnlichkeit stellt eine weitere Schwierigkeit dar: Zwei Wörter können gemäß der besprochenen Methode einander nicht nur ähnlich sein, wenn sie im Text austauschbar sind (z. B. *laufen* und *rennen*), sondern auch, wenn sie jeweils das Gegenteil voneinander sind (z. B. *warm* und *kalt* – die meisten Dinge, die mit *warm* zusammen vorkommen, können auch im Kontext von *kalt* stehen), wenn ein Wort der übergeordnete Begriff des anderen Worts ist (z. B. *Fahrzeug* und *Auto*), oder wenn ein Wort den Teil eines anderen Worts bezeichnet (z. B. *Taste* und *Tastatur*). Es ist durch die beschriebene Methode aber nicht feststellbar, welche Art der Ähnlichkeit vorliegt, sondern nur, dass zwei Wörter aufgrund ihrer Kontexte zusammenhängen.

Ein Beispiel für eine Anwendung, bei der die Methode nach der Distributionellen Hypothese eingesetzt wird, ist in der automatischen Übersetzung die Suche nach der Entsprechung für noch unbekannte Wörter (d. h. neu aufgekomme-

ne Wörter, die noch nicht im angelegten Computer-Wörterbuch stehen). Über den Vergleich der Wortverwendung in ähnlichen Kontexten kann auf das entsprechende Wort in der bekannten Sprache geschlossen werden. Eine weitere Anwendung ist die Internet-Suche, die eine Suchanfrage automatisch um ähnliche Wörter erweitert, um mehr gewünschte, relevante Internetseiten zu finden (beispielsweise wird bei der Eingabe von „Auto" auch nach „Wagen" gesucht).

Ein aktueller Trend ist das Verknüpfen der rein textuellen Wissensquelle mit visuellen Daten – Menschen lernen die Bedeutung von Wörtern schließlich auch nicht exklusiv durch ihre sprachliche Verwendung, sondern auch aus visuellen Informationen. Im Kindesalter ist Kommunikation stark situationsbezogen: In einer der Entwicklungsphasen zeigen Erwachsene auf Gegenstände und benennen sie gleichzeitig – das Kind hört beispielsweise „Apfel", während auf einen fast runden, meist rötlich-gelben Gegenstand gezeigt wird. Andererseits sind auch nicht alle Informationen in Textsammlungen enthalten, weil sie zu offensichtlich sind (Erde ist braun). Bei diesem neuen Ansatz wird ausgenutzt, dass Bilddaten ebenfalls sehr gut mit dem Computer verarbeitbar sind (mit einzelnen Bildpunkten als kleinste Informationseinheit, die größere Bereiche bilden können).

Verstehen Computer also Sprache? – Wann man davon reden kann, dass eine Maschine ‚versteht', ist eine grundsätzliche philosophische Frage. Bei der hier vorgestellten Methode wurde die Bedeutung eines Worts über die Vorkommenshäufigkeiten mit anderen Wörtern als eine Reihe von Zahlen dargestellt, die es ermöglicht hat, seine Bedeutung mit der Bedeutung anderer Wörter zu vergleichen. Der Computer kann damit die Bedeutung eines Worts nicht ausdrücken wie bei einer Definition im Wörterbuch, denn die gewonnene Information ist sehr oberflächlich. Trotzdem kann sie – durch die Möglichkeit, große Textmengen zu verarbeiten – für einen Vergleich mit den gleichartig gewonnenen Bedeutungen anderer Wörter ‚gut genug' dienen und mit der daraus berechneten Bedeutungsähnlichkeit für bestimmte Anwendungen nutzbar gemacht werden. Das heißt, Computer „verstehen" Sprache (noch?) nicht – jedenfalls nicht wie wir Menschen das tun. Aber sie können sie rechnerisch mit intelligenten Methoden verarbeiten – und wir einen Nutzen daraus ziehen.

Zum Abschluss eine verwandte, beliebte Anwendung, die auch auf der Grundlage von Wortvorkommenshäufigkeiten funktioniert: die Wortwolke. Sie bietet eine visuelle Hilfe, die darin unterstützt, die behandelte Thematik einer Textsammlung schnell im Groben zu erschließen. Diese Darstellung besteht aus einer Anordnung der häufigsten Wörter einer Textsammlung, bei der die Größe der Schriftart abhängig von der Anzahl der Vorkommen im Text ist. Häufig genannte Wörter sind also größer und damit prominenter für die Leser. Über die

Farbtonabstufung und Gruppierung der Wörter sind auch weitere Aspekte wie Bedeutungsähnlichkeit für die Betrachter leichter erschließbar zu machen; das ersetzt natürlich nicht die Lektüre des Textes nach geweckten Interesse am Thema. Wie Sie sicher vermuten, wurde das unten abgebildete Beispiel einer Wortwolke aus Textausschnitten der Internetseiten des Instituts für Computerlinguistik in Heidelberg erzeugt.

Literatur

Sahlgren, Magnus: *The word-space model*. Dissertation, Universität Stockholm 2006.

Turney, Peter D. und Pantel, Patrick: *From frequency to meaning: Vector space models of semantics*, in: *Journal of Artificial Intelligence Research* 37 (2010), S. 141–188.