SPRACHE IST MAGIE

Grundlagen der Spracherkennung durch KI

| von JOHN LOUTZENHISER

Indem Sie diesen Artikel lesen, erlauben Sie, liebe Leser, mir, dem Autor, Eindrücke, Ideen und Bilder in Ihrem Kopf entstehen zu lassen. Sie gewähren mir praktisch unbegrenzten Einfluss auf Ihr Inneres, Ihr geistiges Leben – zumindest so lange, wie Sie für diese Lektüre brauchen. Wann immer wir lesen, sprechen oder zuhören, gestehen wir Anderen diese Macht zu – oder üben sie aus. Und die allermeisten von uns halten zu keinem Zeitpunkt inne, um über dieses alltägliche, seltsame und wunderbare Instrument nachzudenken, das uns so viel Macht gibt – die menschliche Sprache. Sprache ist für uns selbstverständlich. Vielleicht, weil Sprache zutiefst menschlich und nur der menschlichen Spezies zu eigen ist. Und weil sie ein so wesentlicher Bestandteil unseres Daseins ist, ist es schwierig Abstand, zu nehmen und sie als etwas von uns Getrenntes zu untersuchen.

Sprache erweist sich bei näherer Betrachtung als schwer durchschaubar und geheimnisvoll. Ihre Eigenschaften sind einzigartig und vielfältig. Als exakter Forschungsgegenstand ist sie daher kaum geeignet.

Doch woher kommt Sprache? Hat der Mensch sie erfunden, wie das Rad oder das Internet? Ist sie ein kulturelles oder ein soziologisches Phänomen? Oder eher ein natürliches, biologisches? Ein Teil von uns, wie das Herz, die Augen oder die Daumen? In welchem Zusammenhang stehen Sprache und Denken? Klären wir unsere Ideen, Absichten und Meinungen für uns selbst bereits in sprachlicher Form, oder denken wir ohne Sprache und nutzen sie erst im zweiten Schritt als Werkzeug, um uns mitzuteilen?

Die Antwort lautet: Sprache ist all dies zusammen: menschliche Erfindung ebenso wie ein natürliches Phänomen. Sie existiert in unserer innersten individuellen Gedankenwelt. Und wir nutzen sie, um mit Anderen, um mit der Außenwelt in Beziehung zu treten.

„DIGITALE UNENDLICHKEIT“

Sprache ist all dies und auch noch etwas gänzlich anderes. Sprache gehört zum Reich der Information. Sie weist eine komplexe Struktur aus Signalen, Symbolen, Regeln und Hierarchien auf – auf den ersten Blick ist die „reine“ Sprache der Logik und Mathematik nicht unähnlich.

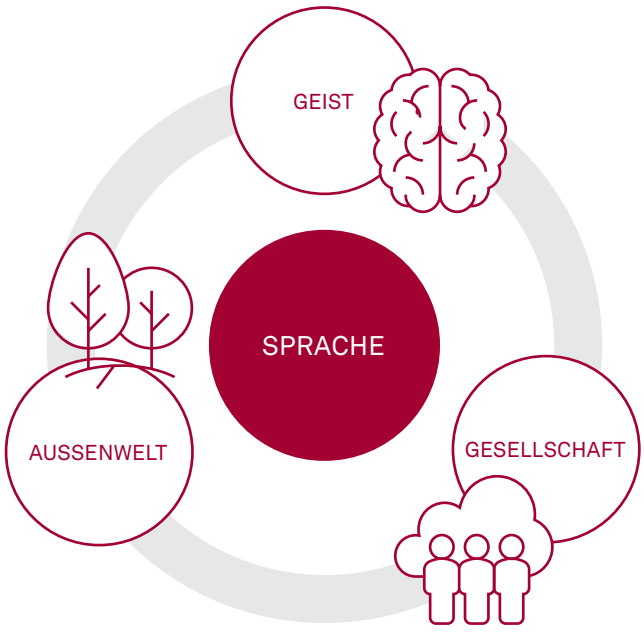

Abbildung 1: Sprache existiert in unserem Kopf, bezieht sich auf die Außenwelt und ist Werkzeug für Kommunikation. Sie ist ein System, das Außenwelt, Gesellschaft und Geist verbindet und von allen drei geformt wird.

> *„Sprache ist im Grunde ein System, das sowohl digital als auch unendlich ist. Soweit ich weiß, gibt es kein anderes biologisches System mit diesen Eigenschaften.“*
> **Noam Chomsky**[1]

Sprache kann eine unendliche Bandbreite unterschiedlichster Erfahrungen darstellen und „kodieren“: unsere Wahrnehmung der äußeren Welt, unsere inneren mentalen, emotionalen und kreativen Erfahrungen, soziale Interaktion und Kommunikation. Dafür verwendet sie einen erstaunlich effizienten und kompakten Code – die Laute und Striche, aus denen gesprochene und geschriebene Sprache besteht. Uns stehen nur wenige Dutzend Schriftzeichen zur Verfügung, um sie zu Worten unseres großen, aber dennoch begrenzten Wortschatzes zusammenzufügen. Wir verbinden diese Worte und formen aus ihnen potenziell unendlich viele verständliche Ausdrücke. So gesehen macht Sprache „unendlichen Gebrauch endlicher Mittel“.

SPRACHE UND KÜNSTLICHE INTELLIGENZ

Diese unbegrenzte expressive Kraft der Sprache und ihre kompakte Darstellungsweise haben sie zu einem fesselnden Untersuchungsgegenstand in der Erforschung künstlicher Intelligenz gemacht – und zwar seit den Anfängen dieser Disziplin vor gut siebzigJahren. Damals wie heute war und ist Sprache der naheliegende, ebenso intuitive wie verführerische Untersuchungsgegenstand: Gelingt es, den Computer „ganz einfach“ zu lehren, die Zeichen und Regeln der Sprache in einer ähnlichen Weise zu handhaben (zu verstehen, einzusetzen), wie wir es tun? Dann wäre er „intelligent“ – oder würde uns zumindest Intelligenz vorgaukeln. Dies ist bekanntlich die Annahme hinter dem Turing-Test[2], einem der frühesten und wichtigsten Impulse in der Geschichte der künstlichen Intelligenz. Einem Computer beizubringen, Sprache zu verwenden, hat sich inzwischen als eine weit größere Herausforderung erwiesen als in den Anfangstagen der KI angenommen. Trotz vieler entscheidender Durchbrüche gibt es auch heute noch keine Software, die Sprache in überzeugender Weise „intelligent“ nutzt, zumindest nicht in dem Sinne, als dass sie den Turing-Test bestehen würde.

Nichtsdestotrotz: Die seit rund zehn Jahren herrschende neue Begeisterung für künstliche Intelligenz geht ganz wesentlich auf jüngere bahnbrechende Fortschritte im Feld der künstlichen Intelligenz beziehungsweise in der „Verarbeitung natürlicher Sprache“, wie es inzwischen heißt (engl.: Natural Language Processing, kurz: NLP), zurück. Diese signifikanten Fortschritte öffnen viele praktische Anwendungsmöglichkeiten für sprachverarbeitende Software.

EINE PRAGMATISCHE WENDE

Die Verarbeitung natürlicher Sprache ist eine Disziplin der Ingenieurswissenschaften und beschäftigt sich mit der Erstellung von „nützlichen" oder in irgendeiner Weise „wertschöpfenden" Artefakten. Dennoch: Ingenieurskunst und Naturwissenschaft gehören eng zusammen. Der Bau von Brücken oder die Raumfahrt beispielsweise erfordern nicht nur einen wissenschaftlichen Hintergrund, sondern bestätigen ihrerseits wissenschaftliche Erkenntnisse und bringen sie weiter voran. Das gilt auch für NLP. Das Bestreben, Software zu entwickeln, die mit einer „intelligenten" Sprachverarbeitung Nutzen und Mehrwerte schafft, hat auch zu Erkenntnissen über das Wesen und die Mechanismen menschlichen Denkens und die menschliche Sprachfähigkeit geführt.

Den Fokus weniger auf theoretische Eleganz zu legen als vielmehr auf den pragmatischen Wert und Nutzen, hat ebenso im Feld der Linguistik und Sprachphilosophie an Bedeutung gewonnen. In weiten Teilen des 20. Jahrhunderts hat sich Linguistik mit formalen, logischen und symbolischen Zugängen zu Syntax (Grammatik) und Semantik (Bedeutung) beschäftigt – in der Hoffnung, eine „reine" und nachvollziehbar adäquate Theorie der Sprache hervorzubringen.

Die Niederungen alltäglichen Sprachgebrauchs, der Einfluss linguistischen und realen Kontextes auf die Bedeutung von Wörtern und Sätzen, der Einfluss von Sprache auf Andere im Verlauf der Kommunikation waren in Randgebiete der Linguistik verbannt – in die Unterdisziplin einer „pragmatischen Linguistik". Heute ist die Pragmatik kein Randgebiet mehr, sondern von zentraler Bedeutung für NLP.

In diesem Artikel stehen zwei der maßgeblichen Ideen der linguistischen Pragmatik und Philosophie der Sprache im Fokus: Kontext und Verwendung.

KONTEXT IST KÖNIG

Kontext ist König: Dieser Gedanke aus der pragmatischen Schule der Linguistik und Sprachphilosophie hat NLP – bezogen auf die Bedeutung von Wörtern und Sätzen – am stärksten geprägt. Er ist im berühmten Zitat von Ludwig Wittgenstein zusammengefasst: „Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache."[3] Der Linguist J. R. Firth formuliert das so: „Die vollständige Bedeutung eines Wortes ist immer kontextabhängig, und keine kontextunabhänige Untersuchungen über Bedeutungen kann ernst genommen werden." Oder noch schärfer: „Man erkennt ein Wort an seinen Begleitern."[4]

Aus Sicht dieser „Kontextualisten" sind Wörter nur im Kontext verständlich. Aber wieso? Bedeuten sie nicht auch an sich etwas, unabhängig vom Zusammenhang? Stimmt es nicht auch, dass Sprache und Wörter auf Dinge in der echten Welt „verweisen" oder „für etwas stehen"? Das Wort „Bedeutung" impliziert genau dieses Sprachverständnis: „auf etwas deuten". Das Wort „Tür" ist in einer befriedigenden Weise ganz einfach ein Symbol, das auf ein Ding in der Welt hinweist, nämlich auf eine Tür. Aber ist das wirklich so einfach? Welche Tür genau? Das lässt sich nur beantworten, wenn es aus dem Zusammenhang heraus, sei er visuell oder verbal, Hinweise auf eine spezifische Tür gibt. Und was genau bedeutet „Tür" im Satz: „Den Bürgermeister getroffen zu haben, hat uns Türen geöffnet"? Ist diese Bedeutung von „Tür" in irgendeiner Weise weniger wirklich als die im Satz „Die Haustür ist offen"?

Offensichtlich gibt es Bezeichnungen mit eindeutigen, feststehenden Bedeutungen. Eigennamen sind ein Beispiel. Es wird sich schwerlich ein Kontext finden, in dem in „München" eine „Brooklyn Bridge" steht. Aber wenn wir Bedeutung ernst nehmen, so wie wir es tun müssen, wenn wir moderne Systeme zur Verarbeitung natürlicher Sprache entwickeln wollen, dann erkennen wir, dass die Bedeutung von Wörtern eine komplexe und unscharfe Angelegenheit ist und sich nicht adäquat über diskrete, also eine begrenzte Anzahl eindeutiger, unterscheidbarer Symbole abbilden lässt. Das Ding, auf das ein Wort vielleicht hinweist, ist nur ein kleiner Aspekt seiner Bedeutung.

Wörter haben komplexe, vieldimensionale Schattierungen und Aspekte. Wir Sprachnutzer spielen damit in kreativer Weise, indem wir eine Teilbedeutung auf neue Situationen anwenden und damit neue, abgeleitete, kontextabhängige Bedeutungen generieren. Deshalb kann „Tür" ebenso „Gelegenheit" oder „Möglichkeit" bedeuten wie das hölzerne Ding mit Angeln (auch so ein Wort) in unserem Haus – immer abhängig vom Kontext.

DIE „EINBETTUNG" ALS DARSTELLUNGSMÖGLICHKEIT VON BEDEUTUNG

Wenn also die Symbole alleine nicht zur Beschreibung der Bedeutung von Wörtern taugen, wie können wir Bedeutung dann fassen, um bessere NLP-Systeme zu entwickeln? Einer der wichtigsten Durchbrüche sind die „Worteinbettungen" (word embeddings), das heißt die kontextabhängige und mehrdimensionale Kodierung von Wörtern. Worteinbettungen ermöglichen einen kontextbezogenen Zugang zu Sprache, so wie oben beschrieben („Sie werden ein Wort an seinen Begleitern erkennen").

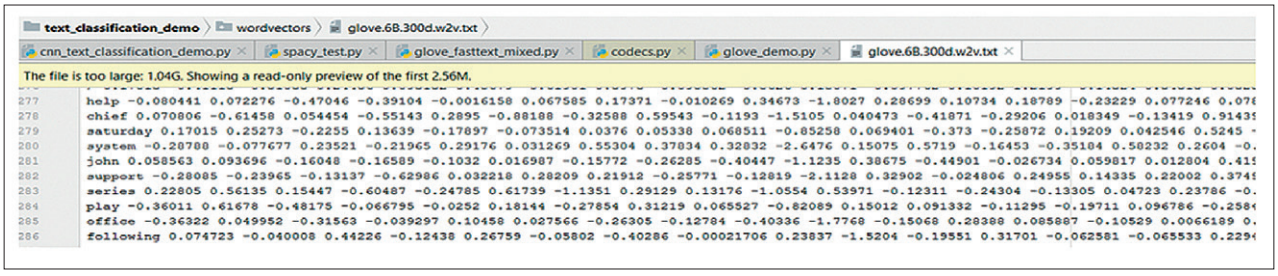

Abbildung 2: Ausschnitt aus einer Worteinbettungsdatei. Jede Zeile ist ein Embedding für ein Wort. Jedes Embedding hat 300 Zahlen (Dimensionen), von denen hier nur die ersten fünfzehn zu sehen sind.

Eine Worteinbettung „kodiert" in gewisser Weise ein Wort inklusive seiner Beziehungen mit und Abhängigkeiten von anderen Wörtern als „Vektor", das heißt als Zahlenfolge aus Dezimalzahlen.

Die Größe des Vektors spiegelt die Anzahl der Dimensionen der Worteinbettung wider. Jede einzelne dieser Dimensionen, so kann man sich das vorstellen, entspricht einem spezifischen Aspekt oder einer Schattierung der Wortbedeutung im Verhältnis zu anderen Wörtern. Im Ergebnis erscheinen Wörter, die in ähnlichen Zusammenhängen auftauchen und die ähnliche Bedeutungen haben, in Gruppen im Vektorraum, zumindest entlang einer oder mehrerer Dimensionen. Eine typische Größe für den Vektor ist in aktuellen KI-Systemen „300" und wird vom Ingenieur festgelegt, der das KI-System trainiert.

PIZZA – ITALIEN + BERLIN = CURRYWURST

Worteinbettungen haben überraschende und nützliche – geradezu magische – Eigenschaften. Nicht nur erscheinen ähnliche Wörter in Gruppen: Wenn man Wortvektoren hinzufügt oder streicht, macht der sich daraus ergebende Vektor oft abstrakte linguistische Ideen oder sogenanntes „Weltwissen" offenbar. Das folgende Beispiel verdeutlicht das:

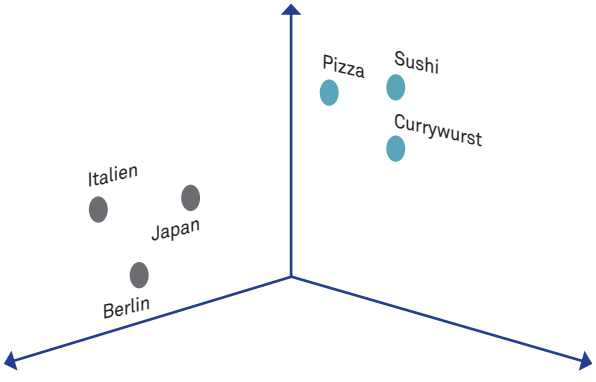

Abbildung 3: Word embeddings – ähnliche Wörter formen „Cluster."

Diese Beispiele sind keine Spielerei, sondern echt. Wenn man eine große Sammlung von Worteinbettungen trainiert, erhält man Vektorrechnungen wie:

vPizza – vItalien + vBerlin = vCurrywurst

Das Ergebnis beruht darauf, dass der Vektor zwischen Pizza und Italien derselbe Vektor ist (mit derselben Richtung und Länge) wie der Vektor zwischen Berlin und Currywurst. Der Vektor scheint also die Vorstellung „typisches Essen" darzustellen.

Ebenso ist der Vektor zwischen Toyota und Japan derselbe wie der zwischen BMW und München. Dieser Vektor scheint also „wichtigster Autohersteller" zu beinhalten.

vToyota – vJapan + vMünchen = vBMW

WORTEINBETTUNGEN TRAINIEREN – ODER „THE MASKED LANGUAGE MODEL"

Natürlich werden Worteinbettungen nicht von Hand eingegeben oder einfach erfunden und niedergeschrieben. Sie sind das Ergebnis von maschinellem Lernen, in der Regel eines neuronalen Netzes, das eine große Sprachsammlung aus Millionen von Sätzen nutzt, beispielsweise alle Einträge in Wikipedia. Der Trainingsprozess lässt sich als eine „Black Box" verstehen, die mit Text gefüttert wird und für jedes einzelne Wort darin einen Wortvektor erstellt.

Bevor man Worteinbettungen mit einem neuronalen Netz trainiert, muss man erst wissen, wofür man es trainiert. Das bedeutet, dass man

1. dem Netz eine bestimmte Aufgabe stellen muss,
2. es mit den für die Lösung dieser Aufgabe relevanten Daten füttern und dann
3. die Fortschritte hin zu einer erfolgreichen Lösung der Aufgabe messen sowie es
4. fortlaufend entsprechend anpassen lassen muss.

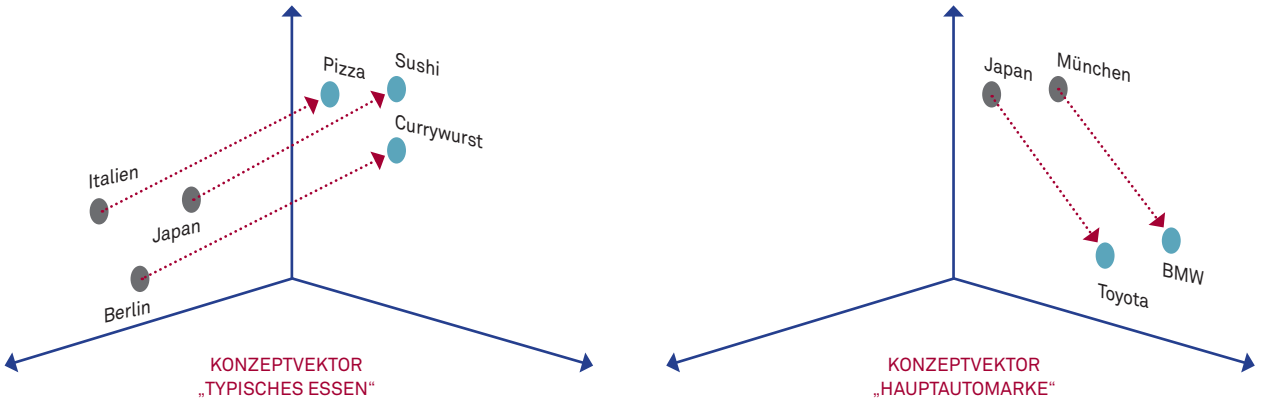

Abbildung 4: Konzeptvektoren in Worteinbettungen

EIN RÄTSEL ZEIGT, WIE ES GEHT

Im Fall des neuronalen Netzes für Worteinbettungen ist die gestellte Aufgabe ein einfaches Rätsel: „Finde das Wort heraus, das im Satz fehlt.“

Dieses Rätsel ist so einfach wie mächtig. Um zu verstehen, wieso diese einfache Aufgabe ausreicht, um anspruchsvolle Worteinbettungen („Sprachmodelle“ oder „Language Model“) zu trainieren, können wir dieses Spiel selber spielen.

Lesen Sie bitte den folgenden Satz:

„Auch im dritten Anlauf hat die britische Regierungschefin Theresa May das mit der Europäischen Union _____ Ausstiegsabkommen nicht über die Ziellinie gebracht.”

Ein Wort fehlt. Um welches Wort handelt es sich? Denken Sie kurz darüber nach, bevor Sie eine Auswahl aus den folgenden Begriffen treffen: a) *abgelehnte* b) *komplizierte* c) *vereinbartem* d) *vereinbarte*

Wir können davon ausgehen, dass alle, die der deutschen Sprache mächtig sind, das gleiche Wort gewählt haben. Welches haben Sie gewählt? Und wieso? Jetzt halten Sie bitte kurz inne und versuchen, Ihre eigene Sprachkompetenz unter die Lupe zu nehmen. Wieso haben Sie das richtige Wort ausgewählt und nicht eines der anderen?

Und jetzt lesen Sie folgenden Satz (Achtung, es ist nicht genau der gleiche Satz wie oben, also aufmerksam lesen!)

„Auch im dritten Anlauf hat die britische Regierungschefin Theresa May das von der Europäischen Union _____ Ausstiegsabkommen nicht über die Ziellinie gebracht.“

Um welches Wort handelt es sich? a) *vorgeschlagene* b) *komplizierte* c) *vereinbartem* d) *vereinbarte*

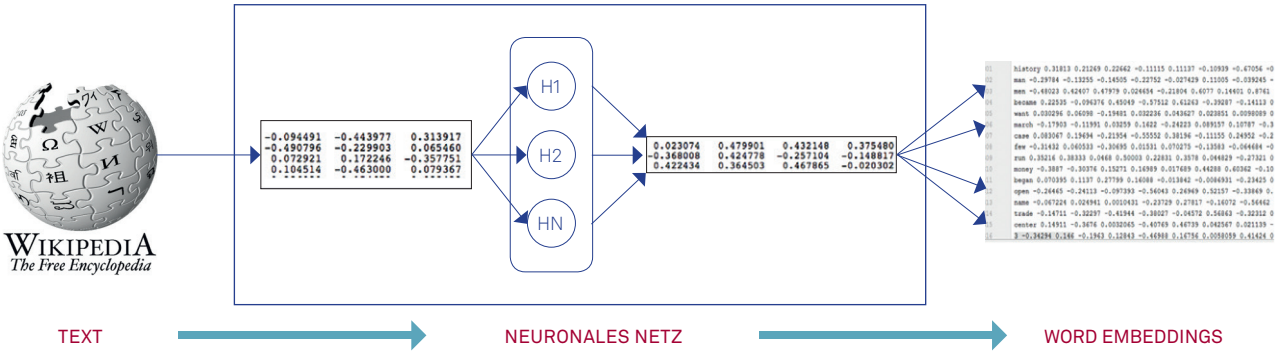

Abbildung 5: Durch ein neuronales Netz erzeugte Worteinbettungen der gesammelten Wikipedia-Texte

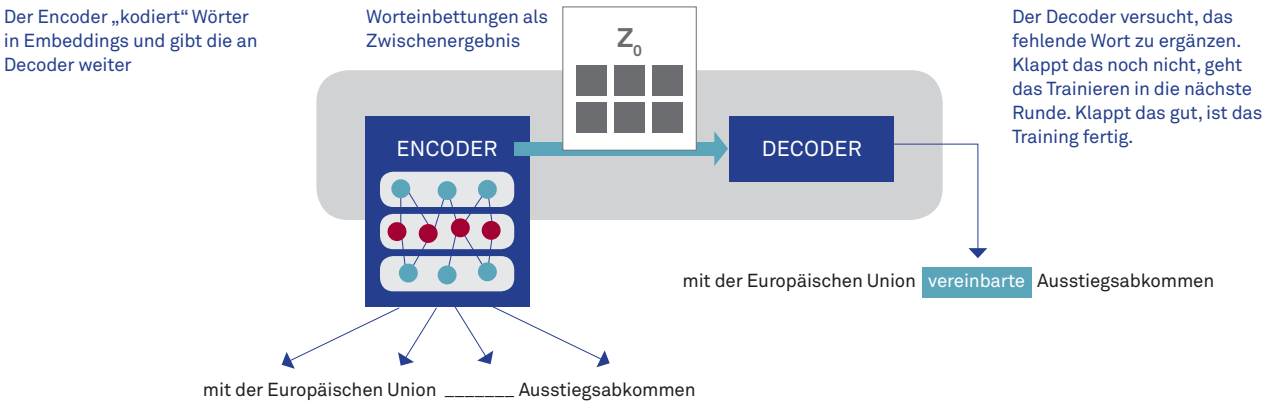

Abbildung 6: Schematische Darstellung des Zusammenspiels von Encoder und Decoder in einem Transformer

Haben Sie das gleiche Wort gewählt? Oder ein anderes? Und wieso? Wir stellen fest, dass wir auf sehr viele Informationen aus dem Gesamtkontext des Satzes gleichzeitig zurückgreifen, um das fehlende Wort richtig zu ergänzen. Weltwissen, Satz-und-Wort-Bedeutung (Semantik), Satzbau und Wortformen (Grammatik, Syntax) – all das fließt in unsere Entscheidung ein, und zwar ohne, dass es uns bewusst ist.

Ebenso, wie wir sehr viel Kontext brauchen, um das fehlende Wort richtig zu ergänzen, ist im Umkehrschluss die Bedeutung von Wörtern extrem kontextabhängig (wie oben bereits erläutert). Worteinbettungen sind dafür die gängige NLP-Methode, und um Worteinbettungen zu erzeugen, trainiert man ein neuronales Netz darauf, genau dieses Rätsel, „Finde das Wort heraus, das im Satz fehlt“, zu lösen.

Um das zu erreichen, benötigt man sehr viele Sätze – in der Regel mehrere Millionen – und „maskiert“ oder „versteckt“ beispielsweise fünfzehn Prozent der Wörter in den Sätzen. Anschließend stellt man einem neuronalen Netz die Aufgabe, diese fehlenden 15 Prozent der Wörter richtig zu ergänzen.

Eine moderne und gängige Netzwerkarchitektur, die dazu verwendet wird, State-of-the-Art-Ergebnisse beim Lösen dieses Rätsels zu erzielen, heißt Transformer. Ein Transformer besteht, einfach ausgedrückt, aus zwei Teilen: einem Encoder und einem Decoder. Der Encoder nimmt den Input – in diesem Fall Trainingssätze mit fünfzehn Prozent versteckten Wörtern – und erzeugt ein Zwischenergebnis, das an den Decoder weitergereicht wird. Der Decoder nimmt das Zwischenergebnis entgegen und versucht, die gestellte Aufgabe zu lösen – in unserem Fall: „Finde das Wort heraus, das im Satz fehlt.“

Das Zwischenergebnis, das der Encoder in diesem Szenario erzeugt, sind genau die oben skizzierten Worteinbettungen (Embeddings). Diese Einbettungen werden im Encoder so lange „trainiert“, bis es dem Decoder gelingt, anhand dieser Einbettungen das richtige Wort zu ergänzen. Sobald der Decoder eine sehr hohe Genauigkeit beim Lösen der Aufgabe aufweist, kann das Training beendet werden. Die Worteinbettungen, die der Encoder erzeugt hat, sind sozusagen ein Nebenprodukt des Trainiervorgangs.

Genauso wie wir Menschen auf unsere sehr stark ausgeprägte Sprachkompetenz zurückgreifen müssen, um das Rätsel des fehlenden Wortes zu lösen, haben auch neuronale Netze, die dieses Rätsel ähnlich gut wie Menschen lösen, eine sehr stark ausgeprägte Sprachkompetenz – oder simulieren sie wenigstens. Man spricht hier von „General Language Understanding“.

Das Elegante an diesem Ansatz ist, dass ein neuronales Netz, das so wie oben beschrieben trainiert wird, mit leichten Anpassungen verwendet werden kann, um eine Vielzahl von NLP-Aufgaben zu lösen – nicht nur die Aufgabe, fehlende Worte zu ergänzen. Solche NLP-Systeme auf Basis von Embeddings können auch für maschinelle Übersetzung, Frage-Antwort-Systeme, automatische Zusammenfassungssysteme und vieles mehr angepasst und verwendet werden. ●

1 „Language is, at its core, a system that is both digital and infinite. To my knowledge, there is no other biological system with these properties.“ Noam Chomsky (1991): Linguistics and Cognitive Science: Problems and Mysteries; in Asa Kasher (ed.), The Chomskyan Turn. Oxford: Blackwell, pp. 26-53, p. 50.
2 https://de.wikipedia.org/wiki/Turing-Test (abgerufen am 11.03.2021).
3 Ludwig Wittgenstein (1953): Philosophische Untersuchungen, §4
4 John Rupert Firth (1957): „A synopsis of linguistic theory 1930-1955“ und (1935) „The Technique of Semantics“